(12)特許協力条約に基づいて公開された国際出願

02 FEB 2005

(19) 世界知的所有権機関
国際事務局

(43) 国際公開日
2004年8月12日 (12.08.2004)

PCT

(10) 国際公開番号
WO 2004/068399 A1

(51) 国際特許分類[7]: G06N 5/00

(21) 国際出願番号: PCT/JP2004/000971

(22) 国際出願日: 2004年1月30日 (30.01.2004)

(25) 国際出願の言語: 日本語

(26) 国際公開の言語: 日本語

(30) 優先権データ:
特願2003-023342 2003年1月31日 (31.01.2003) JP

(71) 出願人 (米国を除く全ての指定国について): 松下電器産業株式会社 (MATSUSHITA ELECTRIC INDUSTRIAL CO., LTD) [JP/JP]; 〒5718501 大阪府門真市大字門真１００６番地 Osaka (JP).

(72) 発明者; および

(75) 発明者/出願人 (米国についてのみ): 森川 幸治 (MORIKAWA, Koji) [/]. 大森 隆司 (OMORI, Takashi) [/]. 大東 優 (OHIGASHI, Yu) [/]. 岡 夏樹 (OKA, Natsuki) [/].

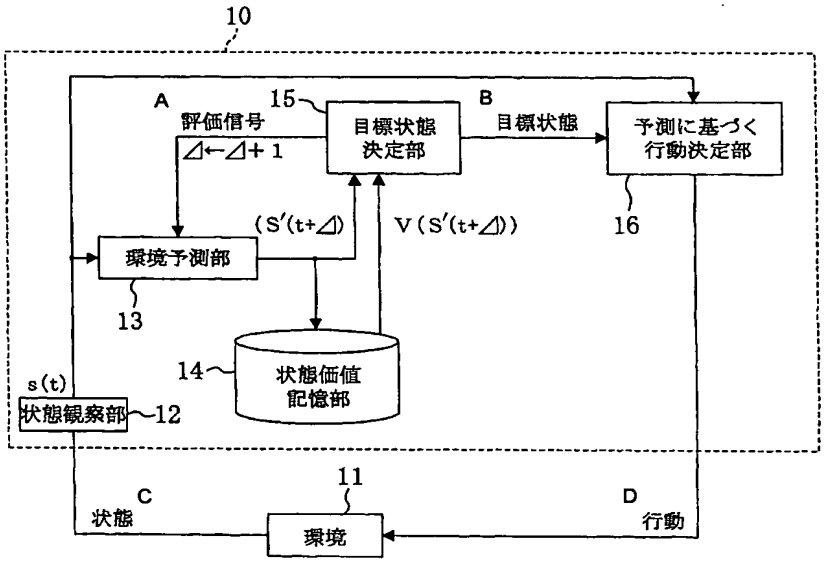
(74) 代理人: 前田 弘, 外 (MAEDA, Hiroshi et al.); 〒5500004 大阪府大阪市靱本町１丁目４番８号 本町中島ビル Osaka (JP).

(81) 指定国 (表示のない限り、全ての種類の国内保護が可能): AE, AG, AL, AM, AT, AU, AZ, BA, BB, BG, BR, BW, BY, BZ, CA, CH, CN, CO, CR, CU, CZ, DE, DK, DM, DZ, EC, EE, EG, ES, FI, GB, GD, GE, GH, GM, HR, HU, ID, IL, IN, IS, JP, KE, KG, KP, KR, KZ, LC, LK, LR, LS, LT, LU, LV, MA, MD, MG, MK, MN, MW, MX, MZ, NA, NI, NO, NZ, OM, PG, PH, PL, PT, RO, RU, SC, SD, SE, SG, SK, SL, SY, TJ, TM, TN, TR, TT, TZ, UA, UG, US, UZ, VC, VN, YU, ZA, ZM, ZW.

[続葉有]

(54) Title: PREDICTIVE ACTION DECISION DEVICE AND ACTION DECISION METHOD

(54) 発明の名称: 予測型行動決定装置および行動決定方法



12...STATE OBSERVATION SECTION
13...ENVIRONMENT PREDICTION SECTION
14...STATE VALUE STORAGE SECTION
15...TARGET STATE DECISION SECTION
A...EVALUATION SIGNAL
B...TARGET STATE
16...ACTION DECISION SECTION BASED ON PREDICTION
C...STATE
D...ACTION
11...ENVIRONEMNT

(57) Abstract: A predictive action decision device (10) includes a state observation section (12) which observes the state of an environment (11) and acquires a state value s(t). An environment prediction section (13) predicts a state change of the environment (11) in future according to the state value s(t). A target state decision section (15) references a state value storage section (14) and decides a future state appropriate for action decision as a target state. An action decision section (16) based on the prediction decides an action according to the target state decided.

(57) 要約: 予測型行動決定装置（１０）において、状態観察部（１２）は環境（１１）の状態を観察し、状態値ｓ（ｔ）を取得する。環境予測部（１３）は状態値ｓ（ｔ）に基づいて、環境（１１）の将来の状態変化を予測する。目標状態決定部（１５）は状態価値記憶部（１４）を参照して、行動決定のために適した将来状態を目標状態として決定する。予測に基づく行動決定部（１６）は決定された目標状態を基にして、行動を決定する。

WO 2004/068399 A1

(84) 指定国 (表示のない限り、全ての種類の広域保護が可能): ARIPO (BW, GH, GM, KE, LS, MW, MZ, SD, SL, SZ, TZ, UG, ZM, ZW), ユーラシア (AM, AZ, BY, KG, KZ, MD, RU, TJ, TM), ヨーロッパ (AT, BE, BG, CH, CY, CZ, DE, DK, EE, ES, FI, FR, GB, GR, HU, IE, IT, LU, MC, NL, PT, RO, SE, SI, SK, TR), OAPI (BF, BJ, CF, CG, CI, CM, GA, GN, GQ, GW, ML, MR, NE, SN, TD, TG).

添付公開書類:
— 国際調査報告書
— 補正書

明　細　書

予測型行動決定装置および行動決定方法

技術分野

本発明は、外部から入力を受け、現在の状態から将来どのような状態に遷移するかを予測しながら、外部への出力を決定する行動決定に関する技術に属する。

背景技術

近年、産業上使用されるシステムは、年々複雑化しており、入力と出力との関係を予めプログラム等によって記述しておくことが困難になりつつある。このため、入力信号を処理して正しい出力を求める方法が必要になっており、このような入力から出力を決定する装置のことを、本願明細書中で「行動決定装置」と呼ぶ。また、特に、入力信号から将来の状態遷移を予測した上で、出力を求めるものを「予測型行動決定装置」と呼ぶ。

行動決定のための従来技術は、１）現在の状態のみによって行動決定を行うもの、２）過去の状態の遷移によって行動決定を行うもの、３）将来の状態を予測して行動決定を行うもの、に分類される。

現在の状態のみによって行動決定を行う従来技術として、ＩＦ－ＴＨＥＮルールを用いるもの、ニューラル・ネットワークを用いるもの、テーブル参照方式などがある。これらは、現在の状態に対する行動が予め記述されており、入力から現在の状態を判別し、所定の記述を参照して行動を決定する。

ところが、現在の状態だけで正しい行動が決定できるとは必ずしも限らない。例えば対話型ロボットシステムにおいて、「いいですか」と聞かれたとき、それだけでは意味不明であり、それまでの状態遷移を参照することによってはじめて言葉の意味を理解することができる。すなわち、行動決定のために、過去の状態が必要となる場合もある。

また、現在や過去の状態だけでなく、将来の状態も考慮すべき場合もある。例

えば、移動ロボットが障害物を回避するような場合には、まだ衝突していない段階では問題は発生していなくて、将来このままの移動方向や速度では衝突するという将来の状態遷移を考慮することによって、はじめて衝突前に回避の行動をとることができる。

将来の状態を考慮した行動決定に関する従来技術として、特許文献１，２に開示された技術がある。特許文献１では、環境から視覚センサや関節角度センサによって得た入力信号から、現在の画像データや関節角度データを状態として受ける。システムは、対象物に対する行動とその行動結果としての画像の変化をリカレント・ニューラル・ネットワークによって記憶し、同じような状況を受けた場合に、記憶している行動を再現する。この技術は、例えばロボットの自律的行動決定に応用されている。また特許文献２では、強化学習における行動決定技術が示されており、ある状態における価値と１ステップ前の状態における価値から、誤差を予測してその情報を行動決定に用いている。

また、特許文献３では、車両の安全走行制御のために、走行経路を推定し、現在の速度で走行を続けた場合に危険な地点がある場合に、その地点に到達する以前に運転者の技量に応じて安全な速度に車両速度を制御する技術が開示されている。

（特許文献１）　特開２００２－５９３８４号公報
（特許文献２）　特開２００２－１８９５０２号公報
（特許文献３）　特開平７－３０６９９８号公報

解決課題

上述のように、特許文献１では、現在状態から自己の行動によってどのような状態に変化するかをリカレント・ニューラル・ネットワークを用いて予測している。そして、その予測結果に応じて、状態と対で記憶された行動が決定される。

しかしながら、特許文献１では、自己の行動に起因する過去の状態変化をリカレント・ニューラル・ネットワークを用いて学習しているに過ぎず、自己の行動と関係のない環境の変化については、予測も考慮も何らなされていない。また、ある時点での行動決定は、現在の状態とその１ステップ後の状態予測に基づいて

行われているが、この１ステップ後の状態が行動決定のために重要であるとは必ずしもいえず、したがって、この将来の状態予測が行動決定にとって適切なものとはいえない。

また、特許文献２に関しても、現在の状態と１ステップ後の状態の予測価値のみから決定された行動が、必ずしも望ましいものとは限らない。例えば、移動ロボットが走ってくる車をよけたいときに、移動ロボットの速度が車に対してかなり遅い場合は、何ステップも前から回避行動を起こさないと車とぶつかってしまうように、将来を見越して行動を決定すべき場合には、１ステップ後だけでなくさらに将来の状態も考慮して行動を決定する必要がある。また、例えば上述の移動ロボットが走ってくる車をよける場合のように、現在や１ステップ後の価値を見ても変化はみられないが、何ステップも先に決定的な自体が発生するときは、現在や１ステップ後の価値に基づく行動は、無駄な行動につながる可能性もある。

また、特許文献３では、地図情報と自車位置情報とを用いて、車が将来進む経路を推定した上で、現在の速度で推定経路を走行した場合に危険な地点が存在するとき、各地点の目標車速を設定した車速計画を作成する。すなわち、車は道路を走行する、という前提の元に、その道路情報を用いて、対処すべき目標となる危険な地点の有無を判断している。ところが、道路情報のような将来状態の情報が予め与えられていない場合には、行動決定を行うための適切な目標を定めること自体が、決して容易ではない。また、この技術では、様々な未経験の事態に対して行動を決定することが、極めて困難である。

発明の開示

本発明は、予測型行動決定装置において、行動決定のために将来の状態予測をより適切に行い、行動決定の精度や能力を向上させることを目的とする。

本発明は、次の点に着目している。すなわち、環境の状態の中には、自己（予測型行動決定装置）の行動（＝出力）に関連するものと、自己の行動に関係なく変化するものとがあり、前者については、長い将来にわたっての状態予測は困難であるが、その一方で、後者については、１ステップ先だけではなく長い将来に

わたっての予測も比較的容易に行うことができる。

すなわち、本発明では、環境を観察して、環境の将来の状態変化を予測し、予測した各将来状態に係る状態価値を参照しつつ、行動決定のために適した将来の状態を、目標状態として決定する。そして、この決定した目標状態を基にして、自己の行動を決定する。これにより、将来状態に係る状態価値を参照して決定された目標状態を基にして、行動が決定されるので、従来よりも行動決定の精度が向上する。

図面の簡単な説明

図１は本発明、および各実施形態を説明するための課題を示す図である。

図２は政策に基づく行動決定に用いられる状態価値と行動基準の例である。

図３は本発明に係る，予測に基づく行動決定を示す図である。

図４は本発明の第１の実施形態に係る予測型行動決定装置の構成を示すブロック図である。

図５は図４の予測型行動決定装置の動作を示すフローチャートである。

図６は本発明の第２の実施形態に係る予測型行動決定装置の構成を示すブロック図である。

図７は図６の構成における価値変換部の動作を示すフローチャートである。

図８は予測ステップ数に伴う状態価値の変化の一例を示すグラフである。

図９は本発明の第３の実施形態における予測に基づく行動決定部の内部構成を示す図である。

図１０は図９の構成の動作を示すフローチャートである。

図１１は本発明の第４の実施形態における環境予測部の内部構成を示す図である。

図１２は図１１の構成の動作を示すフローチャートである。

図１３は本発明の実施形態に係る学習方法と従来の強化学習のシミュレーション結果を示すグラフである。

図１４は本発明の実施形態に係る学習方法と従来の強化学習のシミュレーション結果を示すグラフである。

図１５は本発明の実施形態に係る学習方法と従来の強化学習のシミュレーション結果を示すグラフである。

図１６は状態価値の学習過程を概念的に示す図である。

図１７は学習回数に応じた状態価値曲線の変化を概念的に示す図である。

図１８は本発明の第２の実施形態の変形例に係る予測型行動決定装置の動作を示すフローチャートである。

図１９は本発明の第２の実施形態の変形例における行動選択基準の一例を概念的に示す図である。

発明を実施するための最良の形態

本発明の第１態様によれば、所定の環境について状態を観察し、状態値を取得する状態観察部と、前記環境の各状態に係る状態価値を記憶する状態価値記憶部と、前記状態観察部によって取得された状態値に基づいて、前記環境の将来の状態変化を予測する環境予測部と、前記環境予測部によって予測された各将来状態の中から、前記状態価値記憶部に格納された前記各将来状態の状態価値に基づいて、行動決定のために適した将来の状態を、目標状態として決定する目標状態決定部と、前記目標状態決定部によって決定された目標状態を基にして、自己の行動を決定する第１の行動決定部とを備えた予測型行動決定装置を提供する。

本発明の第２態様によれば、前記環境予測部は、前記環境の将来の自己の行動に影響されない状態変化を予測する第１態様の予測型行動決定装置を提供する。

本発明の第３態様によれば、前記目標状態決定部は、状態価値が極大となる将来状態を目標状態として決定する第３態様の予測型行動決定装置を提供する。

本発明の第４態様によれば、前記状態価値記憶部に記憶された状態価値を学習によって更新する状態価値更新部を備え、前記目標状態決定部は、前記各将来状態の中で、状態価値が前記状態価値更新部によってすでに更新されているものを、前記目標状態として決定する第１態様の予測型行動決定装置を提供する。

本発明の第５態様によれば、前記目標状態決定部は、前記状態価値記憶部から得た状態価値を現在からのステップ数に応じて割り引いて用いる第１態様の予測型行動決定装置を提供する。

本発明の第６態様によれば、前記状態価値記憶部は、自己を含む状態に係る状態価値を記憶するものであり、当該予測型行動決定装置は、前記環境予測部によって予測された，自己を含まない将来状態について、前記状態価値記憶部に記憶された状態価値を基にしてその状態価値を求め、前記目標状態決定部に与える価値変換部を備えた第１態様の予測型行動決定装置を提供する。

本発明の第７態様によれば、所定の行動基準に基づいて自己の行動を決定する第２の行動決定部と、前記第１および第２の行動決定部によって決定された行動を第１および第２の行動候補として受け、これら第１および第２の行動候補のうちのいずれか一方を実際の行動として選択する行動選択部とを備えた第１態様の予測型行動決定装置を提供する。

本発明の第８態様によれば、前記目標状態決定部は、目標状態を決定できたか否かを示す選択信号を前記行動選択部に与えるものであり、前記行動選択部は、前記選択信号が、目標状態を決定できたことを示すときは、前記第１の行動候補を選択する一方、目標状態を決定できなかったことを示すときは、前記第２の行動候補を選択する第７態様の予測型行動決定装置を提供する。

本発明の第９態様によれば、前記第１の行動決定部は、前記状態値を受け、この状態値が表す現在状態からその前ステップにおける状態と行動を検出する行動付状態変化検出部と、前記行動付状態変化検出部によって検出された、現在状態並びにその前ステップにおける状態および行動の組合せを、状態変化として記憶する行動付状態変化記憶部と、前記行動付状態変化記憶部から、現在状態から目標状態までの期間の状態変化の履歴を検索し、この検索結果を基にして、行動を決定する行動計画部とを備えた第１態様の予測型行動決定装置を提供する。

本発明の第１０態様によれば、前記行動計画部は、前記状態変化記憶部の検索の際に、目標状態から現在の状態に向かって後ろ向き探索を行う第９態様の予測型行動決定装置を提供する。

本発明の第１１態様によれば、前記環境予測部は、前記状態値を受け、この状態値が表す現在状態からその前ステップにおける状態を検出する状態変化検出部と、前記状態変化検出部によって検出された現在状態およびその前ステップにおける状態の組合せを状態変化として記憶する状態変化記憶部と、前記状態変化記

憶部から現在状態の後の状態を予測する状態予測部とを備えた第１態様の予測型行動決定装置を提供する。

本発明の第１２態様によれば、予測型行動決定装置において、自己の行動を決定する方法として、所定の環境について状態を観察して、状態値を取得する第１のステップと、取得した状態値に基づいて、前記環境の将来の状態変化を予測する第２のステップと、予測した各将来状態の中から、前記各将来状態の状態価値を参照しつつ、行動決定のために適した将来状態を目標状態として決定する第３のステップと、決定した目標状態を基にして、自己の行動を決定する第４のステップとを備えた行動決定方法を提供する。

本発明の第１３態様によれば、予測する状態変化は、前記環境の将来の自己の行動に影響されない状態変化である第１２態様の行動決定方法を提供する。

本発明の第１４態様によれば、前記第３のステップにおいて、前記各将来状態の中で、状態価値が極大となるものを、前記目標状態として決定する第１２態様の行動決定方法を提供する。

本発明の第１５態様によれば、前記予測型行動決定装置は、前記環境の各状態に係る状態価値を学習によって更新するものであり、前記第３のステップにおいて、前記各将来状態の中で状態価値がすでに更新されているものを、前記目標状態として決定する第１２態様の行動決定方法を提供する。

まず、本発明に関する基本的な概念について説明する。

図１は課題の例を示す図である。図１では、座標（０，０）－（１，１）の空間１において、ボールＢが直進運動をしている。空間１の上下左右の壁に当たったとき、ボールＢは反射する。パドルＰは左右方向にのみ動くことができるものとする。ボールＢの状態は、位置座標（Ｂｘ，Ｂｙ）と進む方向Ｂｔによって表現され、パドルＰの状態は、位置座標（Ｐｘ，Ｐｙ）によって表現される。ただし、Ｐｙ＝０で固定されている。

各時刻ステップにおいて、パドルＰの操作ＭＰとして、｛ＬＥＦＴ（左に動く）、ＲＩＧＨＴ（右に動く）、ＳＴＡＹ（そのまま）｝のうちいずれか１つを選択する。そして、ボールＢをパドルＰによって受けることができたとき、正の報酬が得られるものとし、一方、ボールＢをパドルＰによって受けることができな

かったとき、負の報酬が得られるものとする。そしてここでのタスクは、ボールＢをパドルＰによって受ける回数をできるだけ増やし、累積で得られる報酬をより多くすることである。

これを行動決定問題として見た場合、各ステップごとに、入力としてボールＢおよびパドルＰの状態（Ｂｘ，Ｂｙ，Ｂｔ，Ｐｘ）が与えられたとき、より多くの報酬が得られるように、パドルＰの操作ＭＰを選択する、ということになる。

このような課題に対して、例えば、状態価値および行動基準を用いて、行動を決定する方法がすでに知られている。このような方法を、本願明細書では「政策（Policy）に基づく行動決定」と呼ぶ。

図２は政策に基づく行動決定に用いられる状態価値と行動基準の例を示す図である。ここでの「状態価値」は、外部の状態を評価して得た価値のことである。図２の例では、図１の空間１が８×８のセルに分割されており、各セルには、ボールＢがその位置にあるときの状態価値が示されている。例えばボールＢがセルＣＬの位置にあるとき、状態価値は「３」である。図２の例では、ボールＢが空間１内の下面に到達したときに報酬がもらえるか否かが決定されることから、下面に近い位置ほど状態価値が高くなっている。状態価値は、事前に与えられたり、学習によって獲得されたりする。

また、各セルには、ボールＢがその位置に来たときどの行動を取るべきか、という行動基準が記述されている。図２の例では、図１と対応させて考えると、パドルＰがボールＢの下あたりにあり、ボールＢが左下に向かって落ちてきているときに、ボールＢの位置に対応したセルＣＬにおける行動基準として、左に動く行動に対して０．５、動かない行動に対して０．４、右に動く行動に対して０．１、という値が割り当てられている。この値を基にして、例えば最大値をもつ行動を選択したり、行動選択確率として計算したりして、最終的な行動が決定される。例えば最大値をとる行動を選択する場合には、セルＣＬでは、最大値０．５を持つ「左に動く」が選択される。

このように、政策に基づく行動決定では、現在のボールＢの状態（位置）に対応する状態価値が参照され、その状態に対する行動基準によって行動が決定される。すなわち、行動が現在の状態を考慮して決定されるので、例えばボールＢの

速度が早い場合など、事前に行動を決定する必要があるような場合等には対応できなかった。

これに対して、本発明では、「予測に基づく行動決定」を行う。すなわち図３（ａ）に示すように、現在のボール位置Ｂ１に対して、複数のステップにわたる将来のボール位置を予測し、行動決定に適した目標状態となるボール位置Ｂ２を決定し、この目標状態を基にして行動決定を行う。

図３（ｂ）は予測された各ボール位置に対する状態価値が格納されたテーブル、図３（ｃ）は予測ステップ数に対する状態価値の変化を示すグラフである。図３（ｃ）から分かるように、予測ステップ数を増やすにつれて、ボールＢは下面に向かって移動することになり、これとともに状態価値は徐々に増加する。そして、下面に到達してからはボールＢは上方に向かうため、状態価値は減少に転じる。このため、状態価値は、予測ステップ数５のとき、最大値「８」になる。ここで、例えば最大の状態価値を取った予測ステップ数と予測状態を行動決定に用いるものとすると、位置Ｂ２が目標状態となる。これは、ボールＢの状態でいうと、一番下面に近いときが目標状態として設定されることに相当する。このように、将来の，自己の行動に影響されない状態変化を予測して、予測した状態の中で行動決定のために適したものを目標状態として設定して行動決定を行うことによって、現在の行動をより的確に決定することができる。

以下、本発明の実施の形態について、図面を参照して説明する。なお、以下の説明では、図１の課題をタスクとして扱うものとする。図１の課題は、ピンポンのようにボールが移動してそれを跳ね返すと報酬が得られるものであるが、これは、サッカーゲームにおいてボールを取りに行く行動、移動ロボットが接近物を回避する行動等と同様の設定とみなすことができる。接近物の回避では、接近物の動きを予測してこれを回避することを目標状態として設定することによって、行動を決定できる。この他にも、ロボットが物を受け取る際の行動決定なども、同様の課題となる。

（第１の実施形態）

図４は本発明の第１の実施形態に係る予測型行動決定装置１０の構成を示す。この予測型行動決定装置１０は、所定の環境１１について状態を観察し、環境１

１の，自己の行動に影響されない状態変化を予測した上で、目標状態を定め、環境１１に対する行動を決定する。

状態観察部１２は環境１１の状態を観察し、現在の状態を表す状態値を取得する。ここでは、状態値をｓ（ｔ）と表す。図１の課題については、ボールＢおよびパドルＰの状態（Ｂｘ，Ｂｙ，Ｂｔ，Ｐｘ）が状態値ｓ（ｔ）として得られる。

環境予測部１３は状態観察部１２によって取得された状態値ｓ（ｔ）に基づいて、環境１１の将来の状態変化を予測する。ここでは、ボールＢの座標（Ｂｘ，Ｂｙ）を状態ｓ’として予測するものとする。すなわち、ボールＢは基本的には直進し、壁で反射するので、将来の座標は、現在のボールＢの座標と角度から解析的な計算によって予測することができる。またこの予測には、パドルＰの操作ＭＰを考慮する必要はない。すなわち、当該装置１０の行動に影響されない、将来の状態変化を予測することになる。本実施形態では、環境予測部１３はこのような予測機能を予め備えているものとする。なお、このような予測機能は学習等によって得ることも可能であり、これについては後述する。

状態価値記憶部１４は図３（ｂ）に示すように、各状態ｓ’すなわちボールの位置（Ｂｘ，Ｂｙ）について状態価値をそれぞれ格納している。ここでは、状態ｓ’に対する状態価値をＶ（ｓ’）と記述する。本課題では、ボールＢが下の壁に当たるとき、報酬がもらえるか否かが判定されるので、下の壁に近い位置ほど、より高い状態価値が設定されている。なお、状態価値は全ての状態について設定されているのが好ましいが、これは実際には困難なので、例えば特定の部分状態にのみ状態価値を設定しておいてもよい。

また、状態ｓ’の要素にボールＢの進む方向Ｂｔを含めてもよく、この場合は、状態価値記憶部１４に、（Ｂｘ，Ｂｙ，Ｂｔ）について状態価値をそれぞれ格納しておけばよい。

目標状態決定部１５は、環境予測部１３による予測結果を基にして、行動の決定のために適した将来の状態を目標状態として決定する。ここでは、環境予測部１３によって複数ステップにわたって予測された将来の状態ｓ’（ｔ＋⊿）について、それぞれ、状態価値記憶部１４を参照して状態価値Ｖ（ｓ’（ｔ＋⊿））

を求め、求めた状態価値から目標状態を決定する。

第１の行動決定部としての予測に基づく行動決定部１６は、目標状態決定部１５によって決定された目標状態に対して、現在どのような行動を取るべきかを決定する。ここでは、ボールＢの座標とパドルＰの座標が一致したとき報酬がもらえるので、目標状態から、パドルＰがどの位置に向かえば報酬をもらえる可能性が高くなるを識別し、その位置に向かうように、操作ＭＰを決定する。

以下、図５のフローチャートを参照して、図４の予測型行動決定装置１０の動作を説明する。まず、状態観察部１２によって、環境１１から、現在の状態を表す状態値ｓ（ｔ）が取得される（Ｓ１１）。そして、何ステップ先の予測を行うかを変数⊿で定義し、その初期値として１を与える（Ｓ１２，Ｓ１３）。なお、ステップＳ１７で所定の条件を満たすまで、または、⊿が所定値ｎを超えるまで（Ｓ１９）、⊿をインクリメントしながら、以下のステップＳ１４～Ｓ１６を繰り返し実行する。

ステップＳ１４において、環境予測部１３は、現在状態から⊿ステップ先の状態ｓ’（ｔ＋⊿）を予測する。例えば、ボールＢが壁に当たらないときは、Ｂｘ，Ｂｙが１ステップ分変化し、方向Ｂｔは変化しない。そしてステップＳ１５において、状態価値記憶部１４から、ステップＳ１４で予測された状態ｓ’（ｔ＋⊿）の状態価値Ｖ（ｓ’（ｔ＋⊿））が抽出される。

ステップＳ１６において、目標状態決定部１５は、状態価値記憶部１４から出力された状態価値Ｖ（ｓ’（ｔ＋⊿））を評価する。ここでは、評価条件として、所定の閾値を超えているか否かを判定する。そして、条件を満たす、すなわち状態価値が所定の閾値を越えていると判定したときは（Ｓ１７でＹＥＳ）、ステップＳ１８に進み、状態ｓ’（ｔ＋⊿）を目標状態として、予測に基づく行動決定部１６に与える。

一方、条件を満たさないときは（Ｓ１７でＮＯ）ステップＳ１９に進み、予測ステップ数の判定を行う。すなわち、⊿が所定値ｎを下回るときは（Ｓ１９でＮＯ）ステップＳ１３に戻り、⊿をインクリメントした後、同様の処理を行う。一方、⊿が所定値以上のときは（Ｓ１９でＹＥＳ）ステップＳ１Ａに進み、目標状態が決定できなかった旨を予測に基づく行動決定部１６に通知する。

予測に基づく行動決定部１６は、目標状態決定部１５からの出力を受けて、次の行動を決定する。しかし、目標状態が決定できなかった通知を受けた場合は、例えばランダムに、行動を決定する。

このように本実施形態によると、環境１１の状態変化の予測結果から、状態価値を参照して、行動決定のために適した将来の状態が目標状態として決定される。そして、この目標状態を基にして、自己の行動が決定される。このため、従来の将来を予測した行動決定よりも、行動決定の精度が格段に向上する。また、従来技術のように状態と行動との関係を予め記述していなくても、行動が決定できるので、簡易な構成により、様々な未経験の事態に対しても行動決定可能になる。また、環境１１の将来の，当該装置１０の行動に影響されない状態変化を予測しているので、１ステップ先だけではなく長い将来にわたっての予測も比較的容易に、精度良く、行うことができる。

なお、本実施形態では、状態価値が所定の閾値を越えたときの状態を目標状態として決定するものとしたが、目標状態の決定は、これ以外にも、様々な方法が考えられる。例えば、予測したステップ内で状態価値が最大となった状態や、状態価値が極大になった状態を、目標状態として決定してもよい。また、前ステップとの状態価値の差分が所定値よりも小さくなったときの状態を目標状態として決定してもよい。あるいは、現在からのステップ数に応じて、状態価値を割り引いて評価するようにしてもよい。

なお、本実施形態では、状態価値がテーブル形式で記憶されているものとしたが、この代わりに、ニューラル・ネットワーク等の関数近似手法を適用してもよい。この場合、現在の状態が入力されたとき、状態価値または報酬の期待値が出力されるように学習がなされている必要がある。

（第２の実施形態）

図６は本発明の第２の実施形態に係る予測型行動決定装置１０Ａの構成を示すブロック図である。図６において、図４と共通の構成要素には図４と同一の符号を付している。

まず、価値変換部２１について説明する。価値変換部２１は、状態価値記憶部１４Ａが当該装置１０Ａの行動によって変化する状態を含む状態に係る状態価値

を記憶しているとき、環境予測部１３によって予測された，当該装置１０Ａの行動によって変化する状態を含まない将来状態ｓ’（ｔ＋⊿）の状態価値Ｖ（ｓ’（ｔ＋⊿））を、状態価値記憶部１４Ａに格納された状態価値を基にして求め、目標状態決定部１５Ａに与える。本実施形態では、状態価値記憶部１４Ａは、パドルＰの位置Ｐｘも含めた状態（Ｂｘ，Ｂｙ，Ｂｔ，Ｐｘ）についてそれぞれ状態価値を記憶しているものとし、環境予測部１３からは将来状態（Ｂｘ，Ｂｙ，Ｂｔ）が出力されるものとする。

図７のフローチャートを参照して、価値変換部２１の動作を説明する。まず、環境予測部１３によって将来状態ｓ’（ｔ＋⊿）が予測されたとき、状態価値記憶部１４Ａをこの状態ｓ’（ｔ＋⊿）によって検索する（Ｓ２１）。そして、３個の要素（Ｂｘ，Ｂｙ，Ｂｔ）が一致する状態について、その状態価値の組Ｖ’（ｓ’（ｔ＋⊿））を状態価値記憶部１４Ａから抽出する（Ｓ２２）。ここで、パドルＰの座標Ｐｘが８通りであるとすると、８個の状態価値が探索結果として出力される。

そして、ステップＳ２２において抽出された状態価値を比較し、その最大値を求め（Ｓ２３）、得られた最大値を状態ｓ’（ｔ＋⊿）の状態価値Ｖ（ｓ’（ｔ＋⊿））として目標状態決定部１５Ａに出力する。

また、第１の実施形態では、状態価値記憶部１４に格納された状態価値は事前に与えられるものとしたが、本実施形態では、強化学習と呼ばれる手法によって自動的に学習させるものとする。「強化学習」とは、参考文献１（サットン（R. S. Sutton），バート（A.Barto）著，「強化学習入門（Reinforcement Learning: An Introduction）」，（米国），A Bradford Book, The MIT Press, １９９８年３月）で知られる学習方法の１種であり、環境からの報酬信号と試行錯誤によって学習を行う手法であり、報酬の最大化を目的として行動決定を行う学習アルゴリズムのことである。

本実施形態では、強化学習の中で、ａｃｔｏｒ－ｃｒｉｔｉｃ学習という手法を利用する（参考文献１のｐ．１５１～１５３参照）。ａｃｔｏｒは、どの状態でどの行動を取るべきかを記述する行動決定方策ｐｏｌｉｃｙを持ち、状態ｓで、とりうる行動ａ１～ａｎに対して選択確率が算出されて決定される。ｃｒｉｔ

ｉｃのＶには状態価値と呼ばれる、ある状態でどれぐらい報酬がもらえそうかの期待値を示す値が格納される。状態価値は、報酬から計算されるＴＤ誤差によって更新される。ＴＤ誤差δは、

δ＝ｒ（ｔ＋１）＋γ（Ｖ（ｓ（ｔ＋１））－Ｖ（ｓ（ｔ））

で計算される。ここでｒ（ｔ＋１）は報酬、γは割引率（discount rate）と呼ばれている。このＴＤ誤差によって、ｃｒｉｔｉｃ内部の状態価値テーブルＶ（ｓ）と、ａｃｔｏｒの持つＰｏｌｉｃｙの更新を行う。

ａｃｔｏｒ－ｃｒｉｔｉｃ法を利用して、状態価値記憶部１４Ａに記憶される状態価値を更新することによって、状態価値が不明な状態であっても、予測によって行動を決定することができる。状態価値更新部２２はｃｒｉｔｉｃテーブル更新則によって、状態価値記憶部１４Ａ内に記憶された状態価値を更新する。

ここで、状態価値の学習過程について、図１６を用いて概念的に説明する。状態価値を学習によって獲得する場合、最初は状態価値は与えられておらず、上述した強化学習などの学習法に従って、徐々に状態価値が学習される。

図１６では、図２と同様に、図１の空間１が８×８のセルに分割されており、各セルには、ボールＢがその位置にあるときの状態価値が示されている。同図中、（ａ）は学習前の状態であり、どのセルも状態価値として初期値「０」が与えられている。この場合は、どの位置においても適切な行動は定義されておらず、価値に基づいて行動を決定することはできない。

図１６（ｂ）は学習途中の状態である。状態価値は報酬を受ける毎に更新されていくが、ここではパドルＰがボールＢを打ったときに正の報酬が得られるという設定なので、報酬はボールＢが一番下にあるときにもらえることになる。そして、強化学習によると、状態価値は報酬の伝播のような計算によって算出されるため、図１６（ｂ）に示すように、報酬が得られる位置付近、すなわち下の方だけについて、状態価値が算出されている。この場合は、状態価値が学習されている位置にボールＢが来たときはその状態価値に応じて行動が決定できるが、状態価値が依然として初期値のままの位置では、行動が決定できない。

そして、十分な学習期間が取られた後には、図１６（ｃ）のように、全ての位置に対して状態価値が算出され、どの位置においても状態価値に基づく行動が可

能になる。

また、状態価値の学習過程は、図１７のようにも表すことができる。図１７では、図３（ｃ）と同様に、予測ステップ数に対する状態価値の変化をグラフに表している。ボールＢの将来の位置を予測できる場合は、図１６のような状態価値を参照することによって、横軸に予測ステップ数、縦軸に状態価値をとったときの状態価値の曲線を描くことができる。

上述したように、学習が進むに従って、状態価値はその値を増やし、また報酬がもらえるタイミングｇから何ステップも前の状態に伝播していく。すなわち、学習回数が少ないときは、報酬がもらえるタイミングｇの直前の状態においてのみ状態価値が学習され、学習が進むにつれて、タイミングｇよりもだいぶ前の状態においても状態価値が与えられる。すなわち、状態価値曲線は、学習回数が増加するにつれて、曲線ｃ１，曲線ｃ２，曲線ｃ３のように次第に変化していく。

ここで例えば、曲線ｃ３のように状態価値が学習されたとすると、領域ｅ１は状態価値が割り当てられており、行動が状態価値によって決定できる，いわゆる既学習領域となる。一方、領域ｅ２は、まだ状態価値が学習されておらず、したがって状態価値による行動決定ができない，いわゆる未学習領域となる。曲線ｃ３が図１６（ｂ）の状態に対応しているものと考えると、未学習領域ｅ２はボールＢがまだ上方にある場合に該当する。

学習が十分に進んだ場合は、未学習領域ｅ２はなくなり、既学習領域ｅ１のみになる。これは、例えば図１６（ｃ）のような状態に対応していると考えられる。

このように、学習が進むと、各状態に状態価値が割り当てられていき、状態学習が与えられた既学習領域ｅ１が次第に大きくなるとともに、未学習領域ｅ２が次第に小さくなる。これは、強化学習の特徴の１つともいえる点である。このため、将来状態を複数ステップ予測したときに、状態価値が与えられている既学習領域ｅ１に達するまでにかかるステップ数が、学習が進むにつれて、徐々に少なくなっていくことになる。

図８は強化学習によって学習された状態価値記憶部１４Ａの状態価値に基づき、価値変換部２１が求めた状態価値の変化を示すグラフである。図８において、

縦軸は算出された状態価値、横軸は予測のステップ数である。また図８において、問題設定は図１と同様であるが、状態空間は図２よりも細かく分割されているため、予測ステップ数も多くなっている。図８のグラフでは、予測ステップ数の増加に伴って状態価値が増加しており、８０ステップあたりでピークを迎えている。一般に、状態価値は、図８のように、未学習領域から既学習領域に入ると増加し始め、その後に減少に転じる。この減少に転じるタイミングが、現在状態から予測される最も意義のある状態、すなわち目標状態に相当すると判断できる。

目標状態設定部１５Ａは、第１の実施形態と同様に動作する。ただし、目標状態を決定できたか否かを示す選択信号を、行動選択部２４に与える点が異なっている。予測に基づく行動決定部１６Ａは、第１の実施形態と同様に動作して行動を決定し、決定した行動を第１の行動候補として行動選択部２４に出力する。

第２の行動決定部としての政策に基づく行動決定部２３は、所定の行動基準としての、強化学習におけるａｃｔｏｒ－ｃｒｉｔｉｃ手法によって学習された政策に基づいて行動を決定し、決定した行動を第２の行動候補として行動選択部２４に出力する。

行動選択部２４は、予測に基づく行動決定部１６Ａから受けた第１の行動候補と、政策に基づく行動決定部２３から受けた第２の行動候補とのうち、いずれか一方を、環境１１への実際の行動として選択する。この選択には、目標状態決定部１５Ａから受けた選択信号を利用する。すなわち、選択信号が目標状態を決定できなかったことを示すときは、予測に基づく行動決定部１６Ａから受けた第１の行動候補は意味がないと考えて、政策に基づく行動決定部２３から受けた第２の行動候補を実際の行動として選択する。そうでない場合は、予測に基づく行動決定部１６Ａから受けた第１の行動候補を実際の行動として選択する。

このように本実施形態によると、状態価値を強化学習によって求めることによって、あらかじめ与えることが困難であった状態価値が自律的に獲得され、予測型行動決定が容易に実現できる。また、学習装置としての観点から本装置を見た場合は、報酬が与えられる時点を予測することになり、初期の未学習の状態が多い場合にも行動の指針を得ることができ、学習の効率が向上する。

以下、従来の強化学習（Reinforcement Learning；以下、「ＲＬ」と記載）と、

本実施形態に係る学習方法（Prediction-based Reinforcement Learning；以下「ＰＲＬ」と記載）のシミュレーションによる比較結果をいくつか示す。このシミュレーションでは、ボールが打てたときの報酬は１．０、打てなかったときの報酬はー１．０、またパドルが右または左に動いたときの報酬はー０．０１とした。また、ＰＲＬでは、最初に状態価値記憶部１４Ａ内の状態価値をある程度作成するために３０００試行（ボールが下面に当たるまでを１試行といい、以下epochと記載）は、ＲＬと同様の処理で行動決定を行った。各epochにおけるボールの初期位置はランダムに設定した。

図１３はＲＬとＰＲＬそれぞれの累積報酬の変化を表すグラフである。横軸はepoch数、縦軸は累積報酬である。最初の３０００epochでは、両手法の累積報酬にはほとんど差はない。これは、この期間では、ＰＲＬはＲＬと同じ行動決定方法と学習法で動いているからである。学習の３０００epoch以降では、ＰＲＬの方が、環境予測部１３と予測に基づく行動決定部１６Ａとの利用によって、ＲＬよりも良い性能を示している。これは、ＰＲＬでは、まだ学習が収束していない（すなわち、状態価値記憶部１４Ａ内の状態価値が完全には生成されていない）ときでも、予測によって適切な行動が決定できるので、ボールを打ち返す確率がＲＬよりも高くなるからである。また、ＰＲＬはボールを打ち返す位置の予測に基づき行動を決定しているため、不必要な行動をする割合が低くなっていることもその原因の一つと考えられる。

図１４は学習１００epoch毎に評価を行った結果を示すグラフである。横軸はepoch数、縦軸は１００epochごとに行われた性能評価の結果であり、ボールが打てた割合である。ここでのＲＬの行動決定規則には、ソフトマックス（softmax）と呼ばれる行動戦略を用いている。これは、各状態の各行動に付与された値に応じて確率的に行動決定を行う方法である。初期の３０００epochの学習後、ＰＲＬのパフォーマンスが、ＲＬと比べて大きく向上している。この結果は、まだボールを打ち返す付近の状態しか学習が進んでいない（すなわち、状態価値が決定されていない）ＲＬ学習の初期段階に、ＰＲＬが環境予測部１３と予測に基づく行動決定部１６Ａとの利用によって、学習が進行していない状態に対しても適切な行動を決定できたことが大きな要因である。

図１５はＲＬの行動決定規則をgreedy戦略に変更した場合のグラフである。図１４と比べて、ＲＬの性能は全体的に見ると向上している。しかし、学習していない状態に評価時に遭遇すると性能は急激に落ちる。ＲＬは、ＰＲＬに比べると、精度が悪く安定した動作を示していないことが分かる。

なお、本実施形態では、状態価値をテーブルとして表現して強化学習により学習するものとしたが、ニューラル・ネットワークによって学習させることもできる。この場合、ニューラル・ネットワークの汎化能力によって未経験の状態に対する価値を出力することが期待される。しかし、この手法は、状態価値に不連続性が少ないときに有効であると考えられる。

（第２の実施形態の変形例）

また、学習との組合せを考えた場合、状態価値が未学習の状態から既学習の状態に切り換わるところを目標状態として決定する方法も考えられる。この方法によると、学習がさほど進んでいない状況において、既学習領域では状態価値に基づいて行動決定する一方、未学習領域では、既学習領域に含まれた将来状態を目標状態として決定して、行動決定を行うことができる。

上述の第２の実施形態では、図８に示すグラフにおいて、状態価値が減少に転じるタイミングを目標状態として決定した。ところが、このようにして目標状態を決定できるためには、状態価値が十分な量になるまで蓄積される必要があり、問題によっては、それまでに膨大な学習回数が必要になる。

これに対して、本変形例では、未学習領域から既学習領域に入るタイミングを、目標状態として決定する。これにより、未学習領域が多い学習初期でも、未学習領域では目標状態を基にして行動を決定し、既学習領域では学習によって行動決定することができる。図１７でいうと、未学習領域ｅ２と既学習領域ｅ１とを区別して、行動決定を行うことになる。

本変形例は、図６の予測型行動決定装置１０Ａの構成において、目標状態決定部１５Ａおよび行動選択部２４の動作を変更することによって、実現される。以下、図１８のフローチャートを用いて、その動作を説明する。

まずステップＳ５１において、目標状態決定部１５Ａは、状態観察部１２によって観察された現在の状態Ｓ（例えばボールの現在位置）から、現在の状態価値

Ａを、状態価値記憶部１４Ａおよび価値変換部２１を介して算出する。次にステップＳ５２において、ステップＳ５１で算出された状態価値Ａは、状態価値更新部２２によって既に更新されたことがあるか否か、すなわち既学習か否かを判断する。既学習であるときはステップＳ５３に進み、そうでないときはステップＳ５４に進む。

ステップＳ５３では、現在の状態はすでに学習が進み状態価値が与えられているので、目標状態決定部１５Ａは行動選択部２４に対して、政策に基づく行動決定部２３が決定した第２の行動候補を選択するよう、選択信号を送る。これにより、政策に基づく行動決定部２３によって決定された行動が、環境１１への実際の行動として選択される。

一方、ステップＳ５４では、現在の状態はまだ状態価値が未学習であると考えられるので、環境予測部１３、価値変換部２１および目標状態決定部１５Ａのループ動作によって、将来状態が予測されて、目標状態が決定される。ここでは、予測した将来状態が、既に状態価値が学習されている既学習領域に達したとき、その将来状態を目標状態として決定する。決定された目標状態は、予測に基づく行動決定部１６Ａに送られる。

そしてステップＳ５５において、目標状態決定部１５Ａは行動選択部２４に対して、予測に基づく行動決定部１６Ａが決定した第１の行動候補を選択するよう、選択信号を送る。これにより、予測に基づく行動決定部１６Ａによって決定された行動が、環境１１への実際の行動として選択される。

このように、目標状態決定部１５Ａにおける目標状態の決定基準と、行動選択部２４における行動選択基準とを変更することによって、学習が進んでいない段階であっても、政策に基づく行動決定を行うことができない未学習領域において、予測に基づく行動決定を行うことによって、より広い領域で、行動決定を行うことが可能になる。図１９は本変形例における行動選択を概念的に示した図である。図１９に示すとおり、学習回数が十分でないときでも、予測に基づく行動決定が可能であり、学習が進むにつれて、徐々に、政策に基づく行動決定の割合が高くなっていく。

（第３の実施形態）

上述の実施形態では、予測に基づく行動決定部１６，１６Ａの機能は予め与えられているものとしたが、行動生成機能を予め与えることが難しい場合は、目標状態に到達するための行動生成能力の獲得が必要になる。本実施形態では、このような行動生成能力を学習によって獲得させるものとする。

図９は図６の構成における予測に基づく行動決定部１６Ａの本実施形態に係る内部構成を示す図である。図９において、３１は状態値ｓ（ｔ）を受け、この状態値ｓ（ｔ）が表す現在状態から、その前ステップにおける状態と行動を検出する行動付状態変化検出部、３２は行動付状態変化検出部３１によって検出された現在状態ｓ（ｔ）並びにその前ステップにおける状態ｓ（ｔ－１）および行動ａ（ｔ－１）の組合せを状態変化として記憶する行動付状態変化記憶部、３３は行動付状態変化記憶部３２から、現在状態から目標状態までの期間の状態変化の履歴を検索し、この検索結果を基にして行動を決定する行動計画部である。

図１０のフローチャートを参照して、図９の予測に基づく行動決定部１６Ａの動作を説明する。図１０（ａ）は状態変化を蓄積する場合の動作、図１０（ｂ）は行動計画を行う場合の動作を示す。この２つの動作は、同時並行して実行できる。

状態変化を蓄積する場合、まず、環境１１から状態ｓ（ｔ）を受ける（Ｓ３１）。この現在の状態ｓ（ｔ）とそのときの行動ａ（ｔ）はワーキングメモリに蓄えておく。そして、前ステップにおける状態ｓ（ｔ－１）と行動ａ（ｔ－１）をワーキングメモリから取り出し（Ｓ３２）、状態ｓ（ｔ）とともに行動付状態変化記憶部３２に格納する（Ｓ３３）。これは、状態ｓ（ｔ－１）のときに行動ａ（ｔ－１）をとったら状態ｓ（ｔ）に変化したという、行動に起因する状態変化を表している。

また、行動計画を行う場合、まず、目標状態決定部１５Ａから送られてきた目標状態を探索したい状態ｘｓ（ｎ）として設定する（Ｓ３４）。そして、探索したい状態ｘｓ（ｎ）を行動付状態変化記憶部３２から探索し（Ｓ３５）、検出されたときは（Ｓ３６でＹＥＳ）、探索したい状態ｘｓ（ｎ）と対になって記憶されている状態ｘｓ（ｎ－１）および行動ｘａ（ｎ－１）をワーキングメモリに格納する（Ｓ３７）。その後、ステップＳ３８に進み、探索したい状態の１ステッ

プ前の状態ｘｓ（ｎ－１）が現在の状態ｓ（ｔ－１）でなければ（Ｓ３８でＮＯ）、探索したい状態を更新し（Ｓ３９）、ステップＳ３５に戻る。同様の処理を繰り返し実行し、ステップＳ３８において、探索したい状態の１ステップ前の状態ｘｓ（ｎ－１）が現在の状態ｓ（ｔ－１）と一致したとき（ＹＥＳ）、それまでにワーキングメモリに格納されていた状態ｘｓと行動ｘａの系列を行動計画として出力する（Ｓ３Ｂ）。

一方、ステップＳ３６において、探索したい状態ｘｓ（ｎ）が行動付状態変化記憶部３２から検出できないときは（ＮＯ）、行動不能と判断して（Ｓ３Ａ）、処理を終了する。なおこの場合は、予測に基づく行動決定部１６Ａから、正しい行動が決定できないという信号が出力され、行動選択部２４は、政策に基づく行動決定部２３から出力された第２の行動候補を実際の行動として選択する。

このような動作によって、現在の行動のみならず、現在状態から目標状態に至るまでの行動計画が得られるので、行動計画が一旦完了した後は、その行動計画に従って行動候補を順に出力すればよい。これにより、処理量が格段に少なくなるので、特に長期にわたる予測誤差が少ない場合は、好ましい。もちろん、毎ステップごとに、目標状態までの行動計画を算出し直してもよい。この場合は、予測が完全でない場合であっても、行動が決定できる。

なお、本実施形態では、目標状態から現在状態に向かって後ろ向き探索を行うものとしたが、現在の状態ｓ（ｔ）と行動ａ（ｔ）からｓ（ｔ＋１）を算出する前向き探索を用いても、同様に行動計画を作成することができる。

（第４の実施形態）

本実施形態では、学習によって状態予測を実現するものとする。

図１１は図４および図６の構成における環境予測部１３の本実施形態に係る内部構成を示す図である。図１１において、４１は状態値ｓ（ｔ）を受け、この状態値ｓ（ｔ）が表す現在状態から、その前ステップにおける状態を検出する状態変化検出部、４２は状態変化検出部４１によって検出された現在状態ｓ（ｔ）およびその前ステップにおける状態ｓ（ｔ－１）を状態変化として記憶する状態変化記憶部、４３は状態変化記憶部４２から、現在状態の後の状態を予測する状態予測部である。

図１２のフローチャートを参照して、図１１の環境予測部１３の動作を説明する。図１２（ａ）は状態変化を蓄積する場合の動作、図１２（ｂ）は状態予測を行う場合の動作を示す。この２つの動作は、同時並行して実行できる。

状態変化を蓄積する場合、まず、環境１０から状態ｓ（ｔ）を受ける（Ｓ４１）。この現在の状態ｓ（ｔ）はワーキングメモリに蓄えておく。そして、前ステップにおける状態ｓ（ｔ－１）の組合せをワーキングメモリから取り出し（Ｓ４２）、状態ｓ（ｔ）とともに状態変化記憶部４２に格納する。これは、状態ｓ（ｔ－１）の次に状態ｓ（ｔ）になったという状態変化を表している。

また、状態予測を行う場合、まず、環境１１から取得した現在の状態ｓ（ｔ）を探索したい状態ｙｓ（ｎ）として設定する（Ｓ４４）。そして、探索したい状態ｙｓ（ｎ）を状態変化記憶部４２から探索し（Ｓ４５）、検出されたときは（Ｓ４６でＹＥＳ）、探索したい状態ｙｓ（ｎ）と対になって記憶されている１ステップ後の状態ｙｓ（ｎ＋１）を状態変化記憶部４２から取り出し、出力する（Ｓ４７）。その後、ステップＳ４８に進み、価値変換部２１を経て、目標状態決定部１５，１５Ａから評価信号による再予測依頼を受けたときは（ＹＥＳ）、探索したい状態を更新し（Ｓ４９）、ステップＳ４５に戻る。

一方、ステップＳ４６において、探索したい状態ｙｓ（ｎ）が状態変化記憶部４２から検出できないときは（ＮＯ）、予測先不明と判断して（Ｓ４Ａ）、処理を終了する。なおこの場合は、環境予測部１３から、正しい予測ができないという信号が出力され、行動選択部２４は、政策に基づく行動決定部２３から出力された第２の行動候補を実際の行動として選択する。

このような方法によって、環境予測部１３の機能を予め作成しなくても、学習によって取得することができる。

なお、状態変化記憶部４２の学習のために、ニューラル・ネットワーク等の関数近似手法によって次の状態を予測させることもできる。この場合、ニューラル・ネットワークが本来持つ汎化能力によって、経験したことがない状態ｓ（ｔ）に対しても、適切な１ステップ後の状態ｓ（ｔ＋１）を出力できる可能性がある。

なお、上述の各実施形態では、主に、ピンポンのようなボールを打つ課題を例

に挙げて説明を行ったが、本発明の応用先は、このようなボールを打つ課題に限定されるものではなく、例えば、知能化住宅の制御、情報端末のソフトウェアエージェントの動作決定、ロボットの移動や行動決定など様々な用途が考えられる。

そして、本発明の特徴の１つは、予測した将来状態の中から、状態価値に基づいて、自律的に目標状態を決定することができる点にあり、状態価値が変わると、これに応じて目標状態も自動的に変化し得る。

例えば、知能化住宅においてエアコン等の機器の制御を行い、室内温度を快適にコントロールする場合、ボールの軌跡と同様に、外気温の影響を受けて室内温度がどのように変化するかを予測することは可能である。このとき、将来状態に対する状態価値がすでに学習されている場合（例えば、家の人が帰宅する時刻に近いほど、室内温度が快適になっている方が価値が高い）は、将来状態の中から目標となる時刻と室温の状態を決定することができる。そして、決定した目標状態（帰宅時刻と室温）に向けて、事前に機器の制御を行うことができる。気温の制御等、制御の効果がすぐには現れないような機器制御では、本発明のような予測に基づいた行動決定はきわめて有効である。

また、情報端末でスケジュール管理がなされている場合、ボールの軌跡と同様に、ユーザの将来の予定を参照することができる。このとき、情報提示やサービス提供などの行動（例えば、経済ニュースのダイジェストを提示するサービス）について、将来状態に対して状態価値がすでに学習されている場合（例えば、出張先への移動中が状態価値が高い）、目標状態を決定し、決定した目標状態（移動時刻）に向けて、事前の動作（ニュースの検索、ダウンロード、編集）を開始できる。スケジュールに応じて各将来状態の状態価値が変化すると、それに応じて目標状態も自動的に変化し、適切なタイミングで情報提示やサービス提供を行うことができる。

産業上の利用可能性

本発明では、将来の状態変化がより適切に考慮され、行動決定の精度が向上するので、例えば、知能化住宅の制御、情報端末のソフトウェアエージェントの動

作決定、家庭用ロボットの制御技術などに有効である。

請求の範囲

１．　所定の環境について状態を観察し、状態値を取得する状態観察部と、

前記環境の各状態に係る状態価値を記憶する状態価値記憶部と、

前記状態観察部によって取得された状態値に基づいて、前記環境の将来の状態変化を予測する環境予測部と、

前記環境予測部によって予測された各将来状態の中から、前記状態価値記憶部に格納された前記各将来状態の状態価値に基づいて、行動決定のために適した将来の状態を、目標状態として決定する目標状態決定部と、

前記目標状態決定部によって決定された目標状態を基にして、自己の行動を決定する第１の行動決定部とを備えた

ことを特徴とする予測型行動決定装置。

２．　請求項１において、

前記環境予測部は、前記環境の将来の，自己の行動に影響されない状態変化を予測するものである

ことを特徴とする予測型行動決定装置。

３．　請求項１において、

前記目標状態決定部は、状態価値が極大となる将来状態を、目標状態として決定する

ことを特徴とする予測型行動決定装置。

４．　請求項１において、

前記状態価値記憶部に記憶された状態価値を、学習によって更新する状態価値更新部を備え、

前記目標状態決定部は、前記各将来状態の中で、状態価値が前記状態価値更新部によってすでに更新されているものを、前記目標状態として決定する

ことを特徴とする予測型行動決定装置。

５．　請求項１において、

前記目標状態決定部は、前記状態価値記憶部から得た状態価値を、現在からのステップ数に応じて割り引いて、用いる

ことを特徴とする予測型行動決定装置。

６．　請求項１において、

前記状態価値記憶部は、自己を含む状態に係る状態価値を記憶するものであり、

当該予測型行動決定装置は、

前記環境予測部によって予測された，自己を含まない将来状態について、前記状態価値記憶部に記憶された状態価値を基にしてその状態価値を求め、前記目標状態決定部に与える価値変換部を備えたものである

ことを特徴とする予測型行動決定装置。

７．　請求項１において、

所定の行動基準に基づいて、自己の行動を決定する第２の行動決定部と、

前記第１および第２の行動決定部によって決定された行動を第１および第２の行動候補として受け、これら第１および第２の行動候補のうちのいずれか一方を、実際の行動として選択する行動選択部とを備えた

ことを特徴とする予測型行動決定装置。

８．　請求項７において、

前記目標状態決定部は、目標状態を決定できたか否かを示す選択信号を前記行動選択部に与えるものであり、

前記行動選択部は、前記選択信号が、目標状態を決定できたことを示すときは、前記第１の行動候補を選択する一方、目標状態を決定できなかったことを示すときは、前記第２の行動候補を選択するものである

ことを特徴とする予測型行動決定装置。

９．　請求項１において、

前記第１の行動決定部は、

前記状態値を受け、この状態値が表す現在状態から、その前ステップにおける状態と行動を検出する行動付状態変化検出部と、

前記行動付状態変化検出部によって検出された，現在状態並びにその前ステップにおける状態および行動の組合せを、状態変化として記憶する行動付状態変化記憶部と、

前記行動付状態変化記憶部から、現在状態から目標状態までの期間の状態変化の履歴を検索し、この検索結果を基にして、行動を決定する行動計画部とを備えたものである

ことを特徴とする予測型行動決定装置。

１０．　請求項９において、

前記行動計画部は、前記状態変化記憶部の検索の際に、目標状態から現在の状態に向かって後ろ向き探索を行うものである

ことを特徴とする予測型行動決定装置。

１１．　請求項１において、

前記環境予測部は、

前記状態値を受け、この状態値が表す現在状態から、その前ステップにおける状態を検出する状態変化検出部と、

前記状態変化検出部によって検出された，現在状態およびその前ステップにおける状態の組合せを、状態変化として記憶する状態変化記憶部と、

前記状態変化記憶部から、現在状態の後の状態を予測する状態予測部とを備えたものである

ことを特徴とする予測型行動決定装置。

１２．　予測型行動決定装置において、自己の行動を決定する方法であって、

所定の環境について状態を観察して、状態値を取得する第１のステップと、

取得した状態値に基づいて、前記環境の将来の状態変化を予測する第２のステップと、

予測した各将来状態の中から、前記各将来状態の状態価値を参照しつつ、行動決定のために適した将来状態を、目標状態として決定する第３のステップと、

決定した目標状態を基にして、自己の行動を決定する第４のステップとを備えた

ことを特徴とする行動決定方法。

１３．　請求項１２において、

予測する状態変化は、前記環境の将来の，自己の行動に影響されない状態変化である

ことを特徴とする行動決定方法。

１４．　請求項１２において、

前記第３のステップにおいて、

前記各将来状態の中で、状態価値が極大となるものを、前記目標状態として決定する

ことを特徴とする行動決定方法。

１５．　請求項１２において、

前記予測型行動決定装置は、前記環境の各状態に係る状態価値を、学習によって更新するものであり、

前記第３のステップにおいて、

前記各将来状態の中で、状態価値がすでに更新されているものを、前記目標状態として決定する

ことを特徴とする行動決定方法。

補正書の請求の範囲

［２００４年6月０３日（０３．０６．０４）国際事務局受理：出願当初の請求の範囲２及び１３は取り下げられた；出願当初の請求の範囲１，５，１１及び１２は補正された；他の請求の範囲は変更なし。（５頁）］

１．（補正後）　所定の環境について状態を観察し、状態値を取得する状態観察部と、

前記環境の各状態に係る状態価値を記憶する状態価値記憶部と、

前記状態観察部によって取得された状態値に基づいて、前記環境の将来の状態変化を予測する環境予測部と、

前記環境予測部によって予測された各将来状態の中から、前記状態価値記憶部に格納された前記各将来状態の状態価値に基づいて、行動決定のために適した将来の状態を、目標状態として決定する目標状態決定部と、

前記目標状態決定部によって決定された目標状態を基にして、自己の行動を決定する第１の行動決定部とを備え、

前記環境予測部は、前記環境の将来の，自己の行動に影響されない状態変化を予測するものである

ことを特徴とする予測型行動決定装置。

２．（削除）

３．　請求項１において、

前記目標状態決定部は、状態価値が極大となる将来状態を、目標状態として決定する

ことを特徴とする予測型行動決定装置。

４．　請求項１において、

前記状態価値記憶部に記憶された状態価値を、学習によって更新する状態価値更新部を備え、

前記目標状態決定部は、前記各将来状態の中で、状態価値が前記状態価値更新部によってすでに更新されているものを、前記目標状態として決定する

ことを特徴とする予測型行動決定装置。

５．（補正後）　所定の環境について状態を観察し、状態値を取得する状態観察部と、

前記環境の各状態に係る状態価値を記憶する状態価値記憶部と、

前記状態観察部によって取得された状態値に基づいて、前記環境の将来の状態変化を予測する環境予測部と、

前記環境予測部によって予測された各将来状態の中から、前記状態価値記憶部に格納された前記各将来状態の状態価値に基づいて、行動決定のために適した将来の状態を、目標状態として決定する目標状態決定部と、

前記目標状態決定部によって決定された目標状態を基にして、自己の行動を決定する第１の行動決定部とを備え、

前記目標状態決定部は、前記状態価値記憶部から得た状態価値を、現在からのステップ数に応じて割り引いて、用いる

ことを特徴とする予測型行動決定装置。

６．　請求項１において、

前記状態価値記憶部は、自己を含む状態に係る状態価値を記憶するものであり、

当該予測型行動決定装置は、

前記環境予測部によって予測された，自己を含まない将来状態について、前記状態価値記憶部に記憶された状態価値を基にしてその状態価値を求め、前記目標状態決定部に与える価値変換部を備えたものである

ことを特徴とする予測型行動決定装置。

７．　請求項１において、

所定の行動基準に基づいて、自己の行動を決定する第２の行動決定部と、

前記第１および第２の行動決定部によって決定された行動を第１および第２の行動候補として受け、これら第１および第２の行動候補のうちのいずれか一方を、実際の行動として選択する行動選択部とを備えた

ことを特徴とする予測型行動決定装置。

８．　請求項７において、

前記目標状態決定部は、目標状態を決定できたか否かを示す選択信号を前記行動選択部に与えるものであり、

前記行動選択部は、前記選択信号が、目標状態を決定できたことを示すときは、前記第１の行動候補を選択する一方、目標状態を決定できなかったことを示すときは、前記第２の行動候補を選択するものである

ことを特徴とする予測型行動決定装置。

９．　請求項１において、

前記第１の行動決定部は、

前記状態値を受け、この状態値が表す現在状態から、その前ステップにおける状態と行動を検出する行動付状態変化検出部と、

前記行動付状態変化検出部によって検出された，現在状態並びにその前ステップにおける状態および行動の組合せを、状態変化として記憶する行動付状態変化記憶部と、

前記行動付状態変化記憶部から、現在状態から目標状態までの期間の状態変化の履歴を検索し、この検索結果を基にして、行動を決定する行動計画部とを備えたものである

ことを特徴とする予測型行動決定装置。

１０．　請求項９において、

前記行動計画部は、前記状態変化記憶部の検索の際に、目標状態から現在の状態に向かって後ろ向き探索を行うものである

ことを特徴とする予測型行動決定装置。

１１．（補正後）　所定の環境について状態を観察し、状態値を取得する状態観察部と、

前記環境の各状態に係る状態価値を記憶する状態価値記憶部と、

前記状態観察部によって取得された状態値に基づいて、前記環境の将来の状態変化を予測する環境予測部と、

前記環境予測部によって予測された各将来状態の中から、前記状態価値記憶部に格納された前記各将来状態の状態価値に基づいて、行動決定のために適した将来の状態を、目標状態として決定する目標状態決定部と、

前記目標状態決定部によって決定された目標状態を基にして、自己の行動を決定する第１の行動決定部とを備え、

前記環境予測部は、

前記状態値を受け、この状態値が表す現在状態から、その前ステップにおける状態を検出する状態変化検出部と、

前記状態変化検出部によって検出された，現在状態およびその前ステップにおける状態の組合せを、状態変化として記憶する状態変化記憶部と、

前記状態変化記憶部から、現在状態の後の状態を予測する状態予測部とを備えたものである

ことを特徴とする予測型行動決定装置。

１２．（補正後）　予測型行動決定装置において、自己の行動を決定する方法であって、

所定の環境について状態を観察して、状態値を取得する第１のステップと、

取得した状態値に基づいて、前記環境の将来の状態変化を予測する第２のステップと、

予測した各将来状態の中から、前記各将来状態の状態価値を参照しつつ、行動決定のために適した将来状態を、目標状態として決定する第３のステップと、

決定した目標状態を基にして、自己の行動を決定する第４のステップとを備え、

予測する状態変化は、前記環境の将来の，自己の行動に影響されない状態変化である

ことを特徴とする行動決定方法。

１３．（削除）

１４．　請求項１２において、

前記第３のステップにおいて、

前記各将来状態の中で、状態価値が極大となるものを、前記目標状態として決定する

ことを特徴とする行動決定方法。

１５．　請求項１２において、

前記予測型行動決定装置は、前記環境の各状態に係る状態価値を、学習によって更新するものであり、

前記第３のステップにおいて、

前記各将来状態の中で、状態価値がすでに更新されているものを、前記目標状態として決定する

ことを特徴とする行動決定方法。

FIG. 1

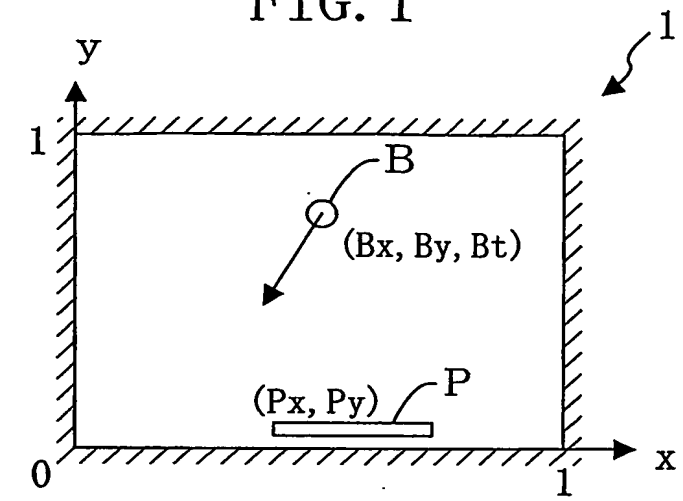

y
1
1
B
(Bx, By, Bt)
(Px, Py)
P
0
1
x
Mp(LEFT, RIGHT, STAY)

FIG. 2

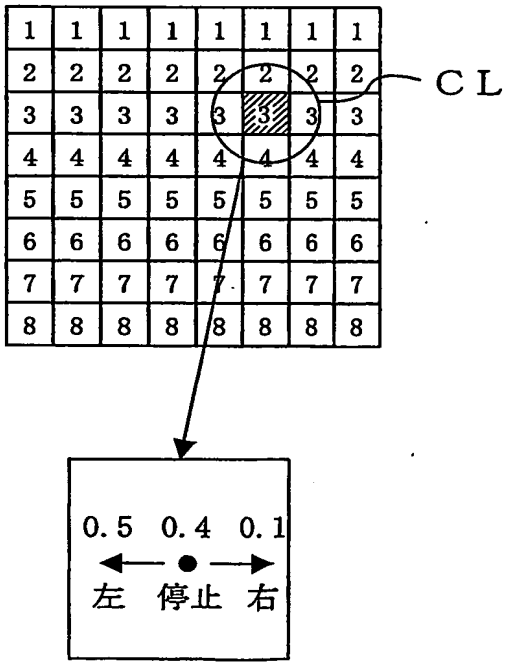

1 1 1 1 1 1 1 1
2 2 2 2 2 2 2 2
3 3 3 3 3 3 3 3
4 4 4 4 4 4 4 4
5 5 5 5 5 5 5 5
6 6 6 6 6 6 6 6
7 7 7 7 7 7 7 7
8 8 8 8 8 8 8 8
CL
0.5 0.4 0.1
左 停止 右

FIG. 3

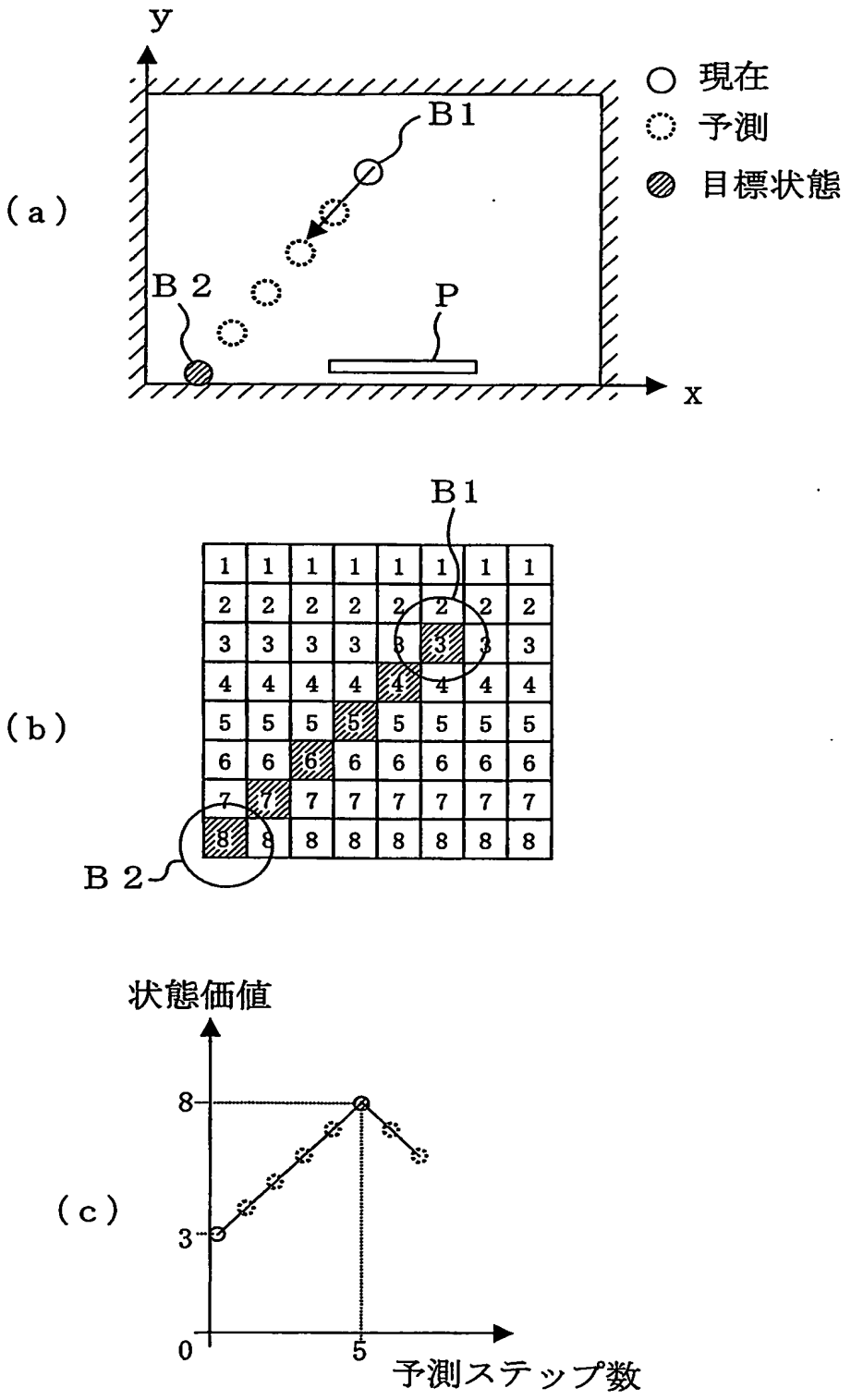


FIG. 4

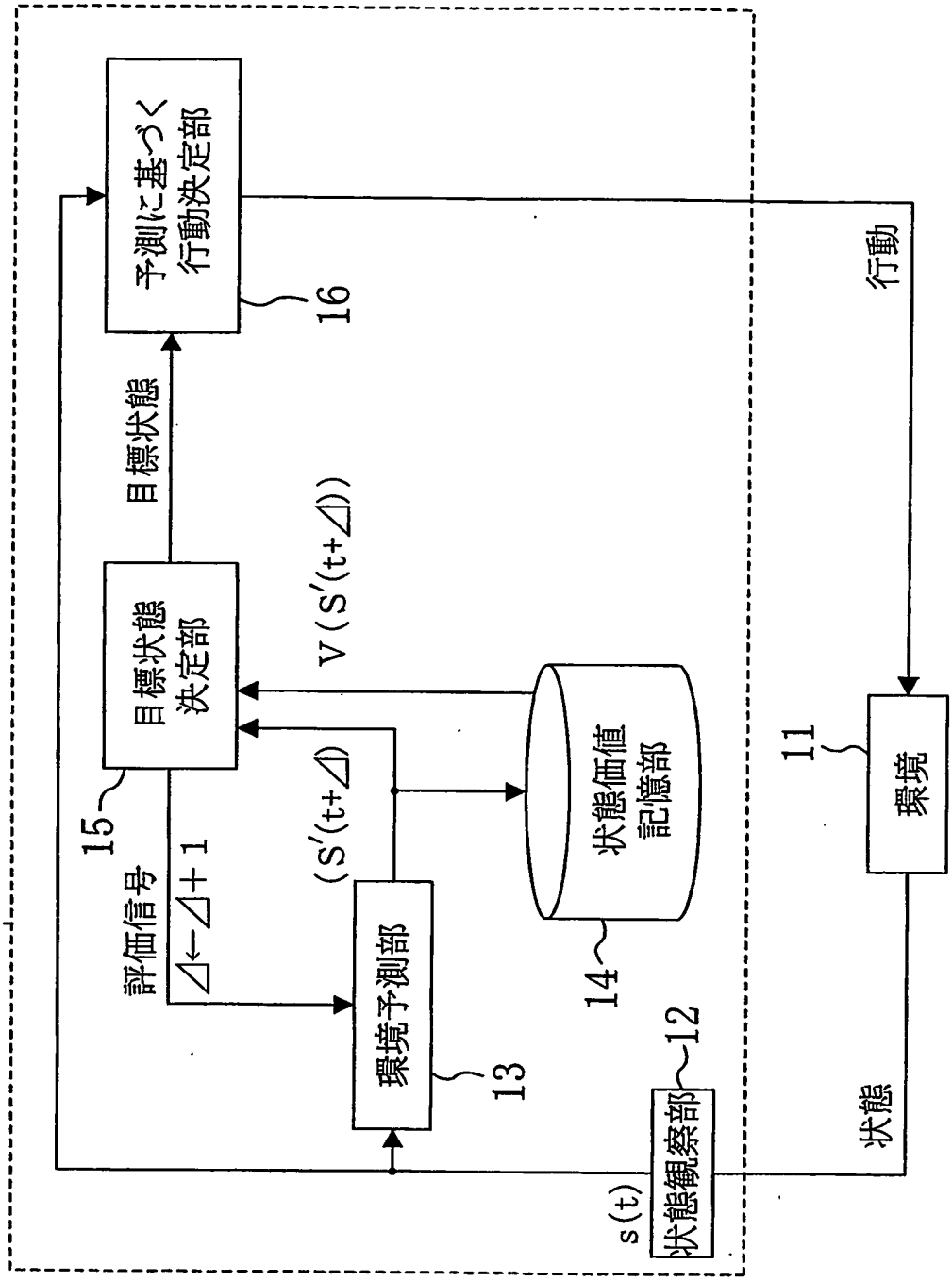

10
予測に基づく行動決定部
16
目標状態
目標状態決定部
15
評価信号
Δ←Δ+1
V(S'(t+Δ))
(S'(t+Δ)
環境予測部
13
状態価値記憶部
14
状態観察部
12
s(t)
11
環境
行動
状態

FIG. 5

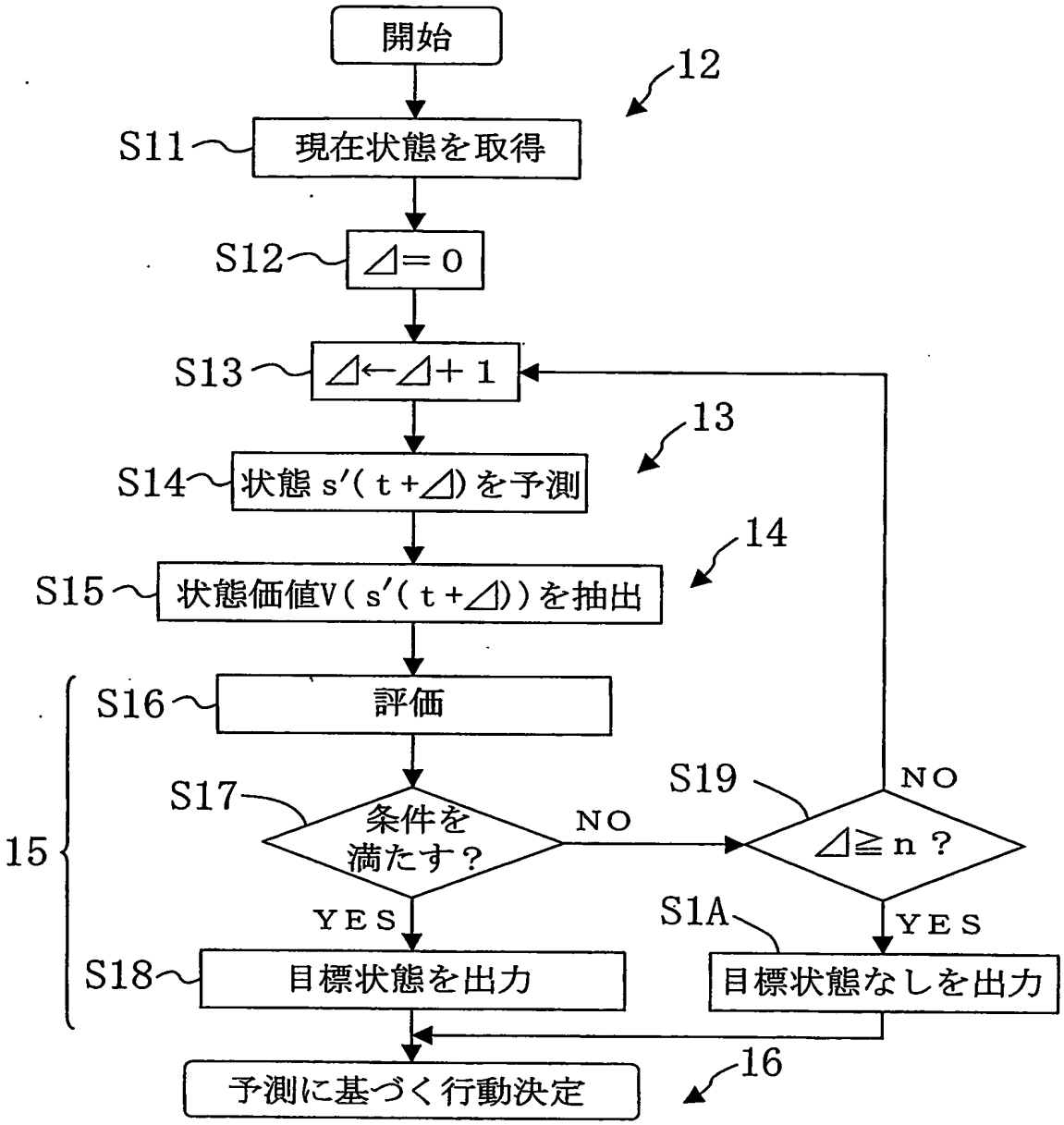

開始
12
S11
現在状態を取得
S12
⊿=0
S13
⊿←⊿+1
13
S14
状態 s'(t+⊿)を予測
14
S15
状態価値V(s'(t+⊿))を抽出
S16
評価
S17
条件を満たす？
NO
S19
⊿≧n？
NO
15
YES
YES
S1A
S18
目標状態を出力
目標状態なしを出力
16
予測に基づく行動決定

FIG. 6

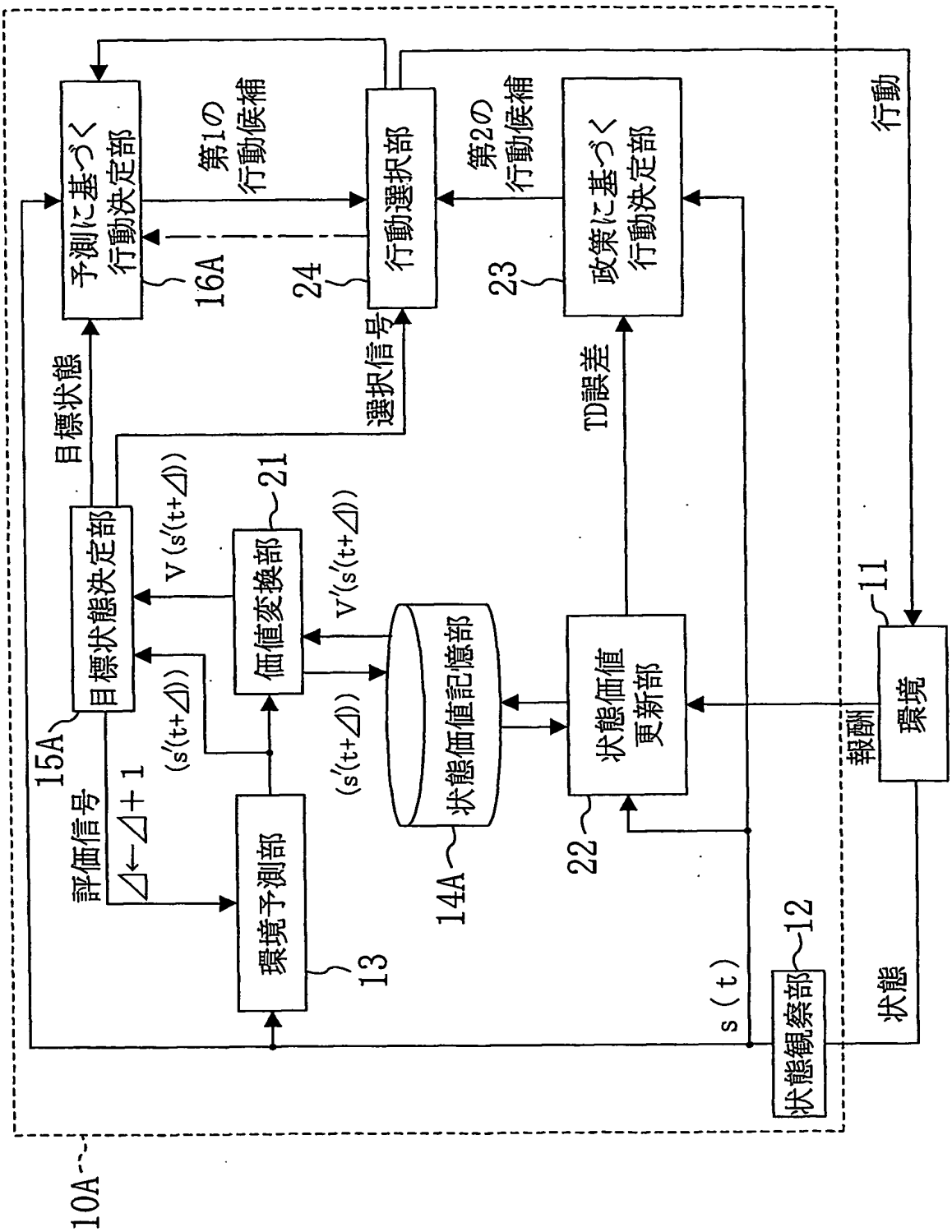

10A
予測に基づく行動決定部
16A
第1の行動候補
行動選択部
24
第2の行動候補
政策に基づく行動決定部
23
行動
目標状態
選択信号
TD誤差
目標状態決定部
15A
V(s'(t+Δ))
価値変換部
21
V'(s'(t+Δ))
状態価値記憶部
14A
状態価値更新部
22
11
環境
報酬
評価信号
Δ←Δ+1
(s'(t+Δ))
(s'(t+Δ))
環境予測部
13
s(t)
状態観察部
12
状態

FIG. 7

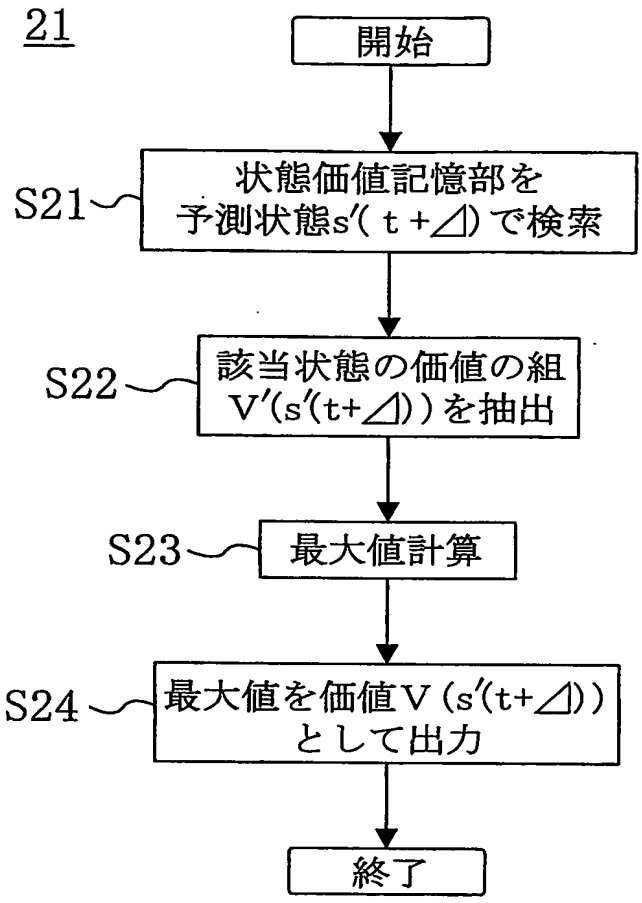

21
開始
S21
状態価値記憶部を
予測状態s′(t+⊿)で検索
S22
該当状態の価値の組
V′(s′(t+⊿))を抽出
S23
最大値計算
S24
最大値を価値V(s′(t+⊿))
として出力
終了

FIG. 8

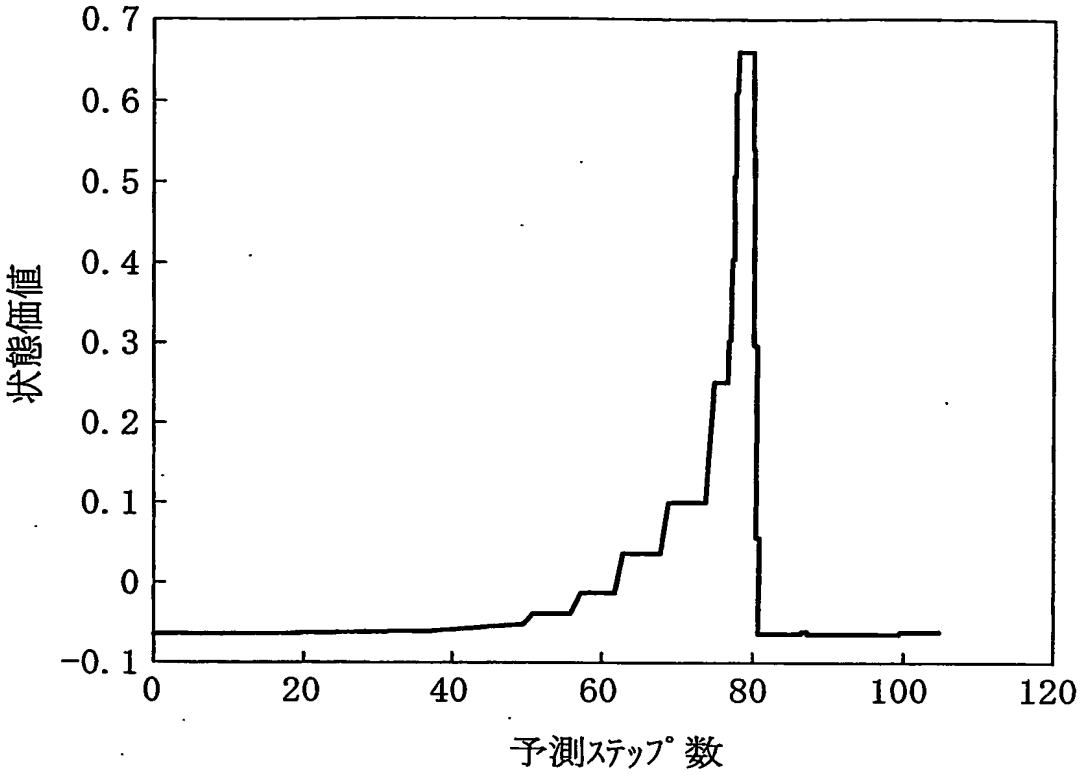

0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
-0.1
状態価値
0
20
40
60
80
100
120
予測ステップ数

FIG. 9

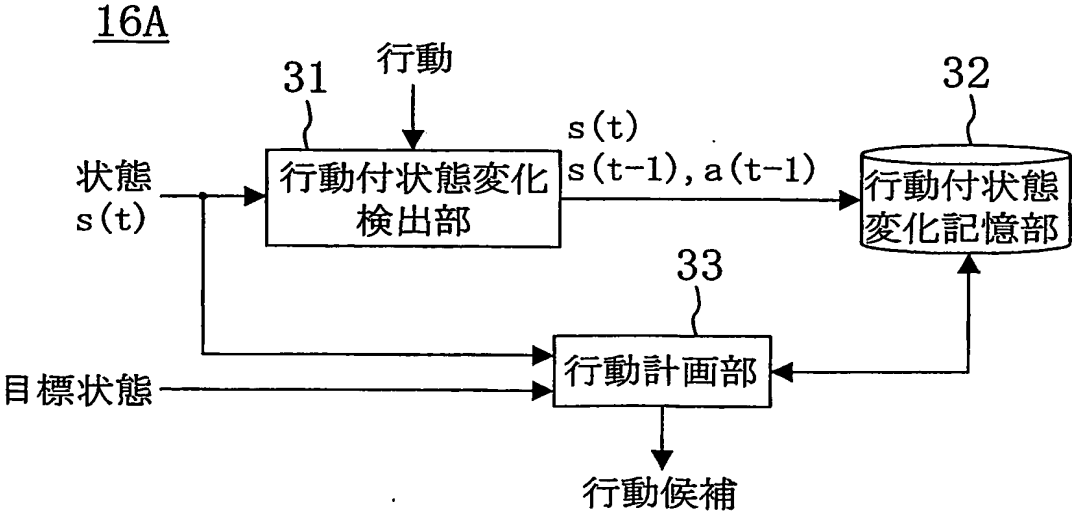

16A
31
行動
32
s(t)
状態
s(t)
行動付状態変化
検出部
s(t-1), a(t-1)
行動付状態
変化記憶部
33
目標状態
行動計画部
行動候補

FIG. 10

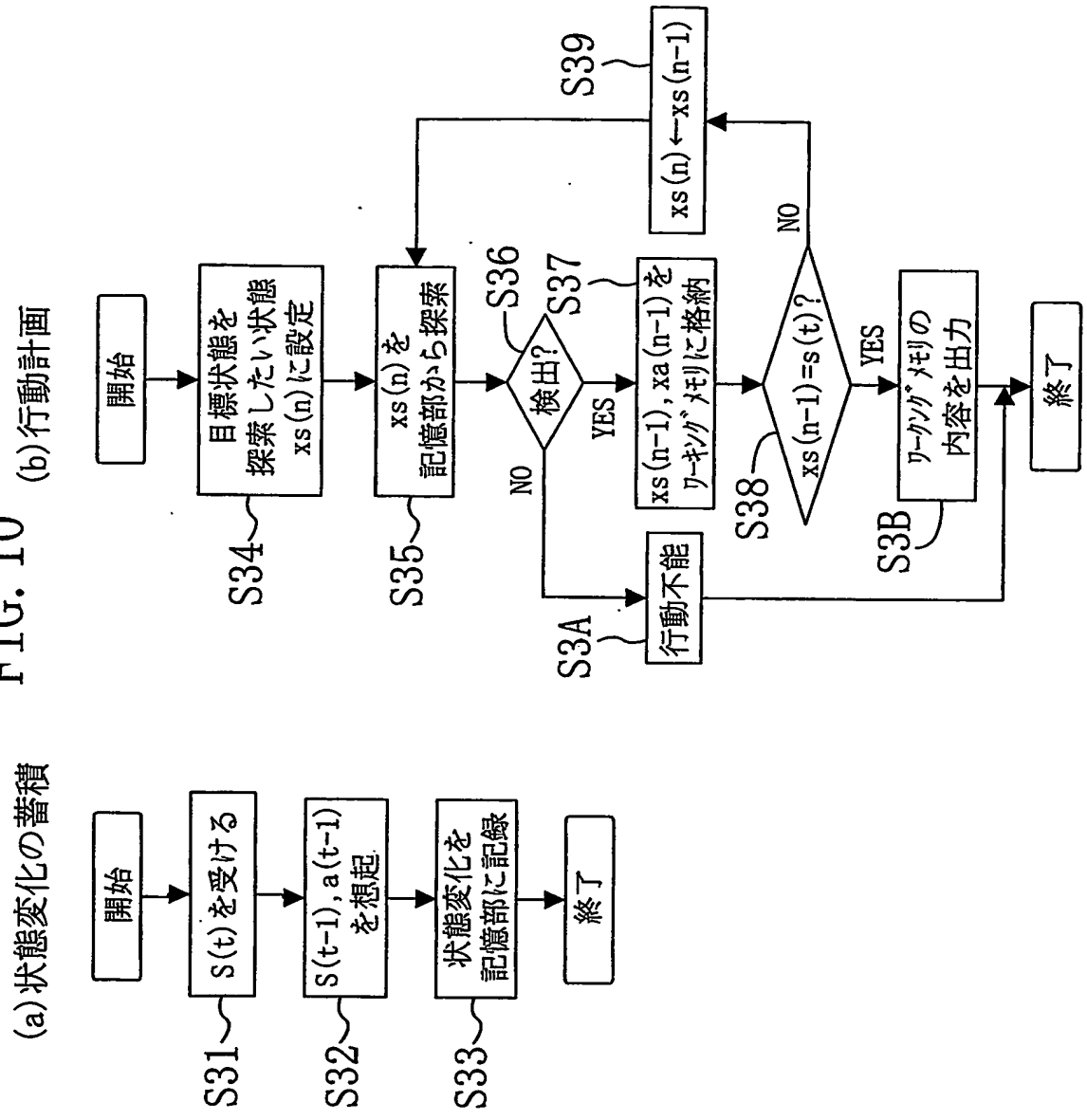

(a)状態変化の蓄積
開始
S(t)を受ける
S(t-1), a(t-1)を想起
状態変化を記憶部に記録
終了
S31
S32
S33
(b)行動計画
開始
目標状態を探索したい状態xs(n)に設定
xs(n)を記憶部から探索
検出?
YES
NO
xs(n-1), xa(n-1)をワーキングメモリに格納
xs(n-1)=s(t)?
YES
NO
xs(n)←xs(n-1)
行動不能
ワーキングメモリの内容を出力
終了
S34
S35
S36
S37
S38
S39
S3A
S3B

FIG. 11

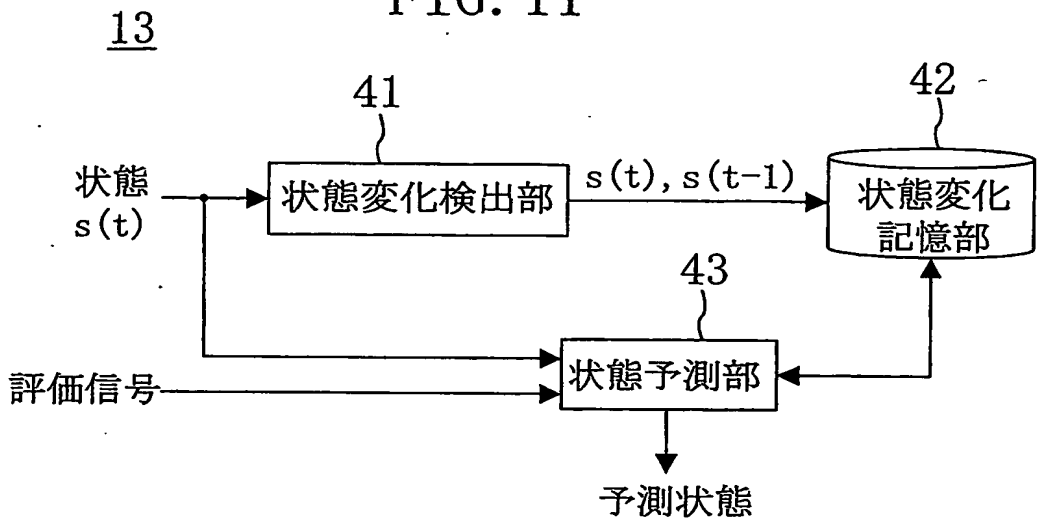

13
41
42
状態
s(t)
状態変化検出部
s(t), s(t-1)
状態変化
記憶部
43
評価信号
状態予測部
予測状態

FIG. 12

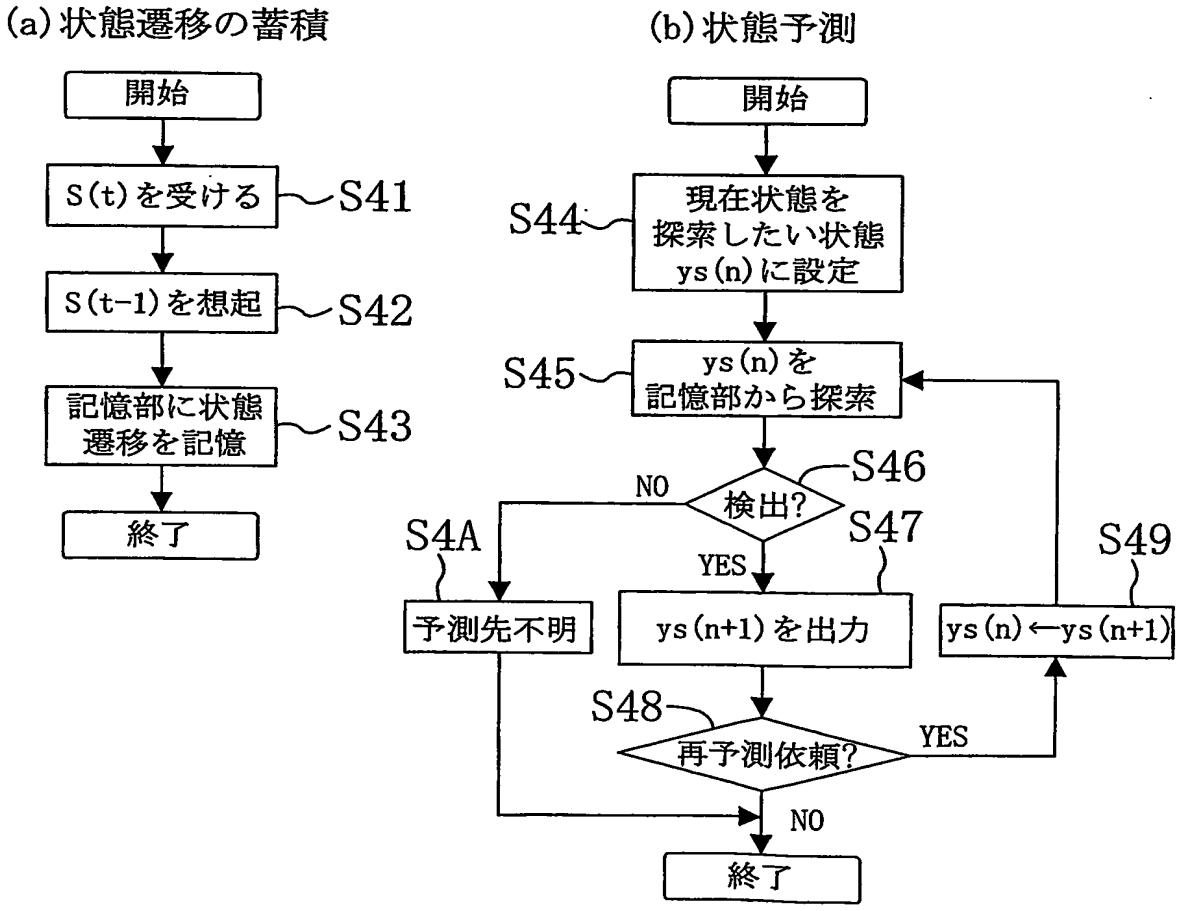

(a) 状態遷移の蓄積
開始
S(t)を受ける
S41
S(t-1)を想起
S42
記憶部に状態
遷移を記憶
S43
終了
(b) 状態予測
開始
S44
現在状態を
探索したい状態
ys(n)に設定
S45
ys(n)を
記憶部から探索
S46
NO
検出?
YES
S4A
S47
S49
予測先不明
ys(n+1)を出力
ys(n)←ys(n+1)
S48
再予測依頼?
YES
NO
終了

FIG. 13

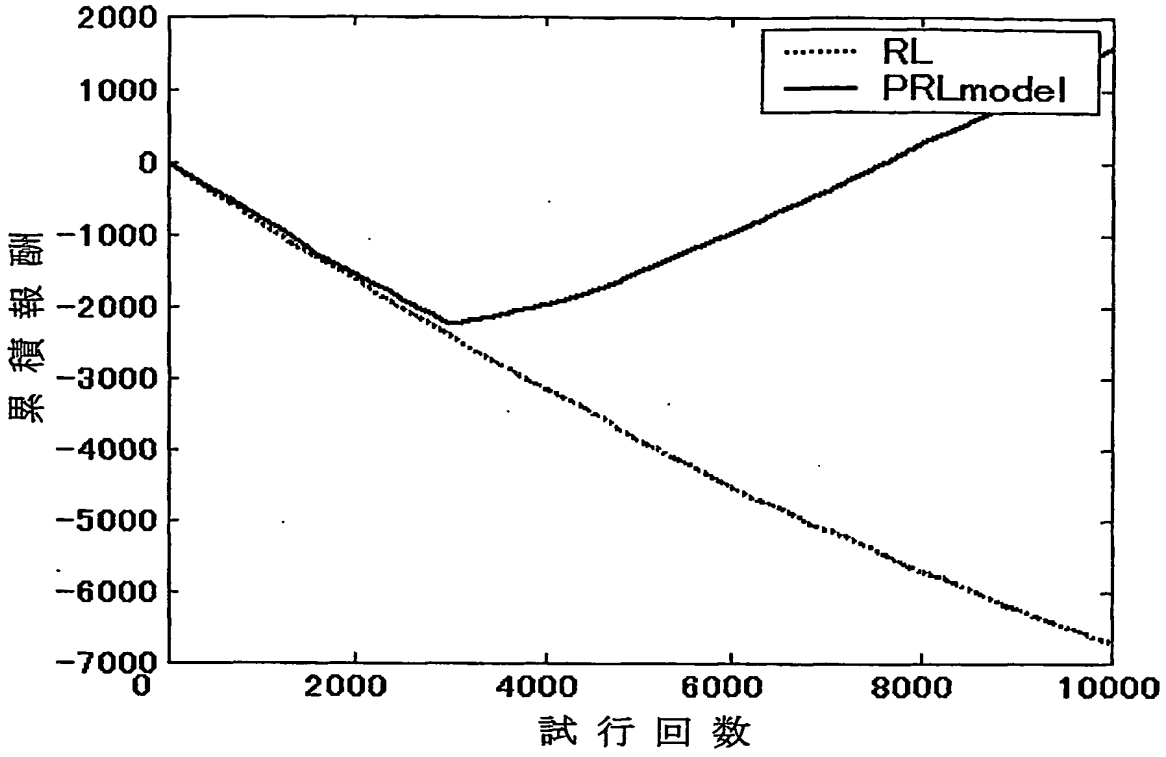

RL
PRLmodel
累積報酬
2000
1000
0
−1000
−2000
−3000
−4000
−5000
−6000
−7000
0
2000
4000
6000
8000
10000
試行回数

FIG. 14

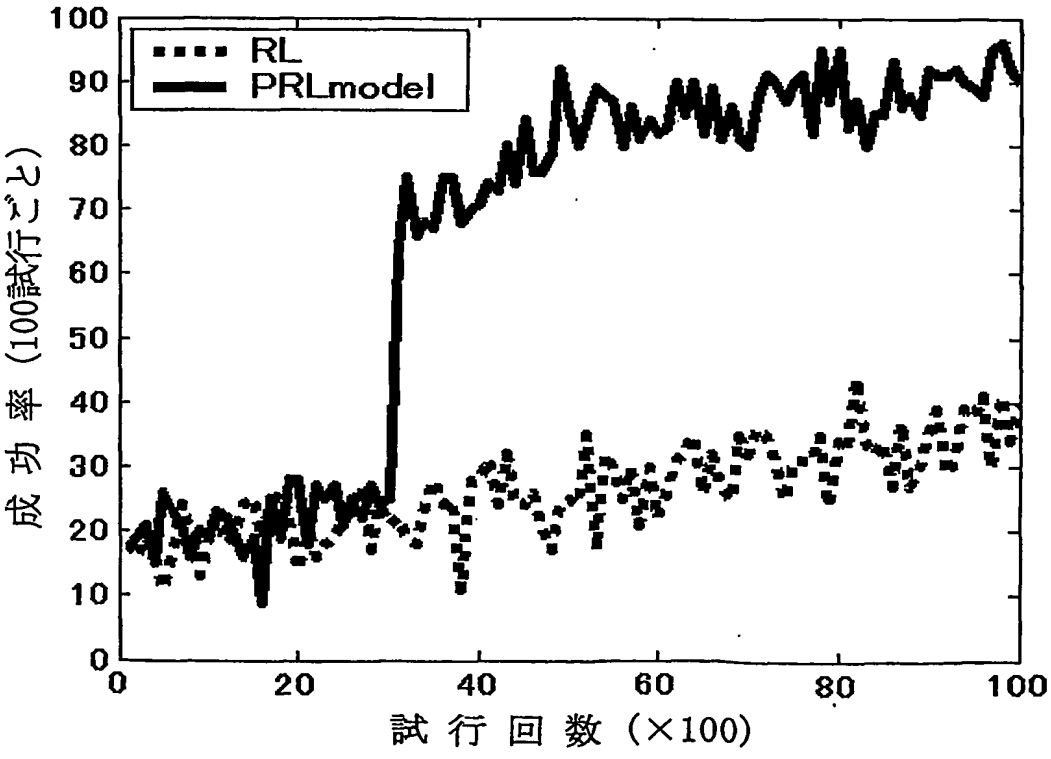

RL
PRLmodel
成功率（100試行ごと）
100
90
80
70
60
50
40
30
20
10
0
0
20
40
60
80
100
試行回数（×100）

FIG. 15

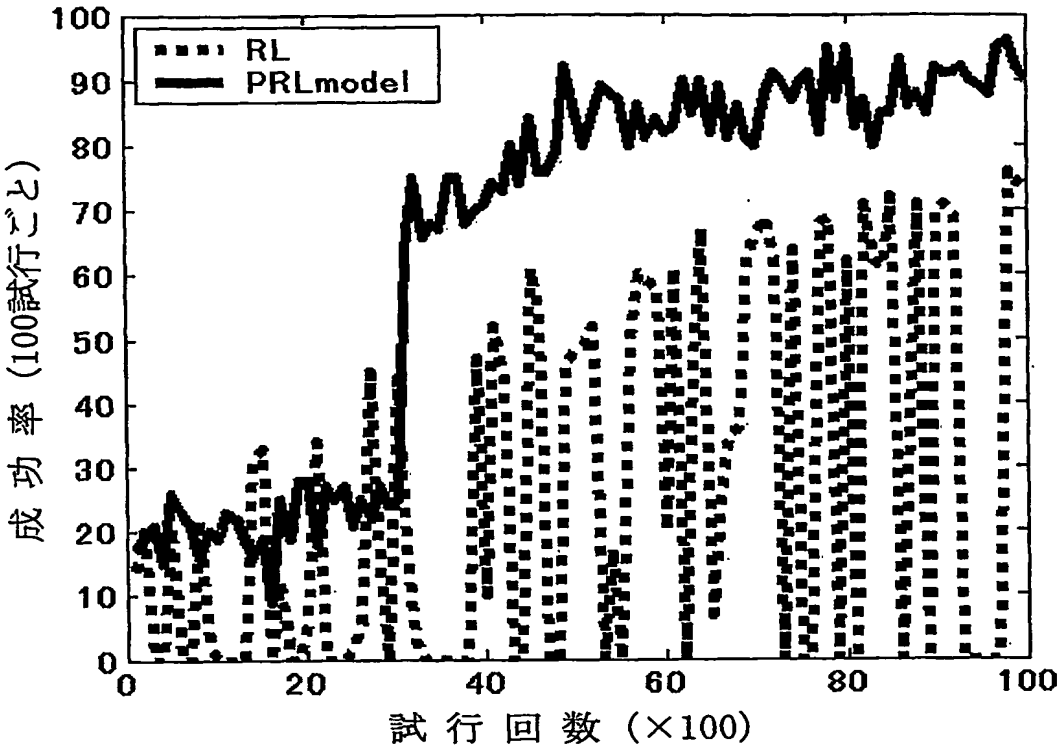

RL
PRLmodel
成功率（100試行ごと）
試行回数（×100）
0
10
20
30
40
50
60
70
80
90
100
0
20
40
60
80
100

FIG. 16

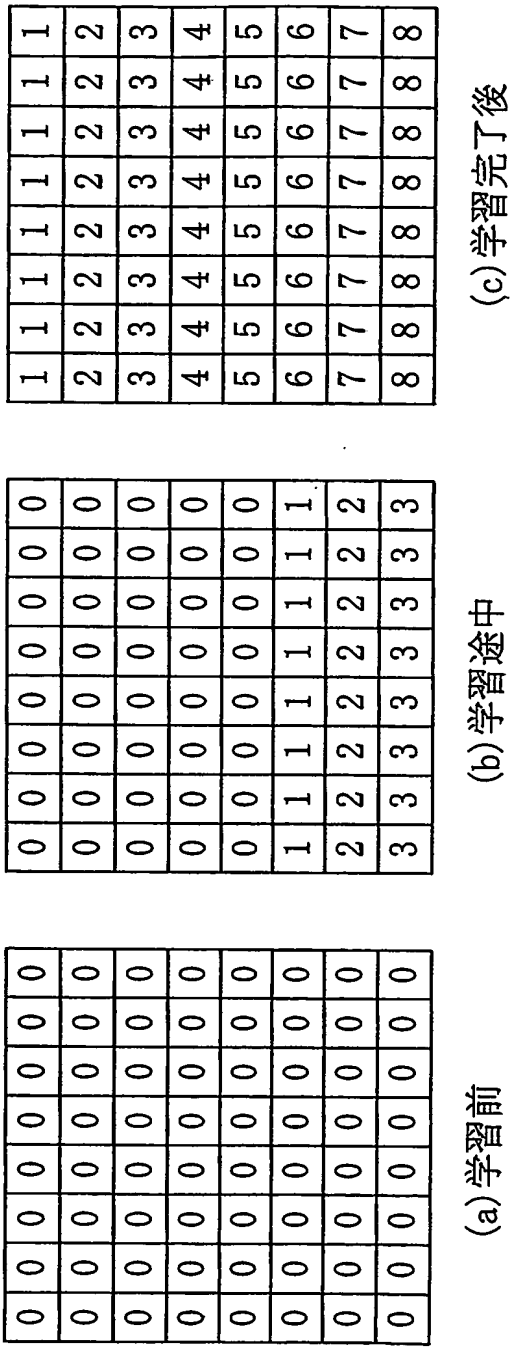


FIG. 17

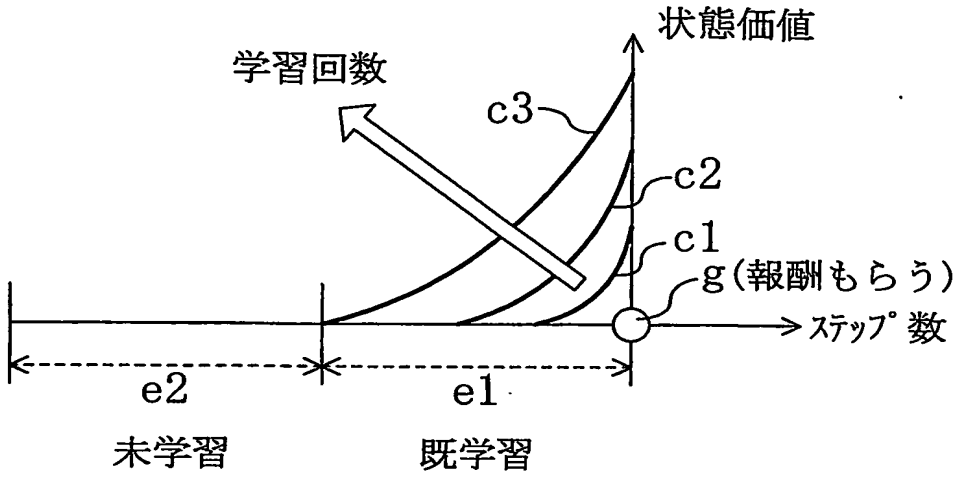

状態価値
学習回数
c3
c2
c1
g(報酬もらう)
ステップ数
e2
e1
未学習
既学習

FIG. 18

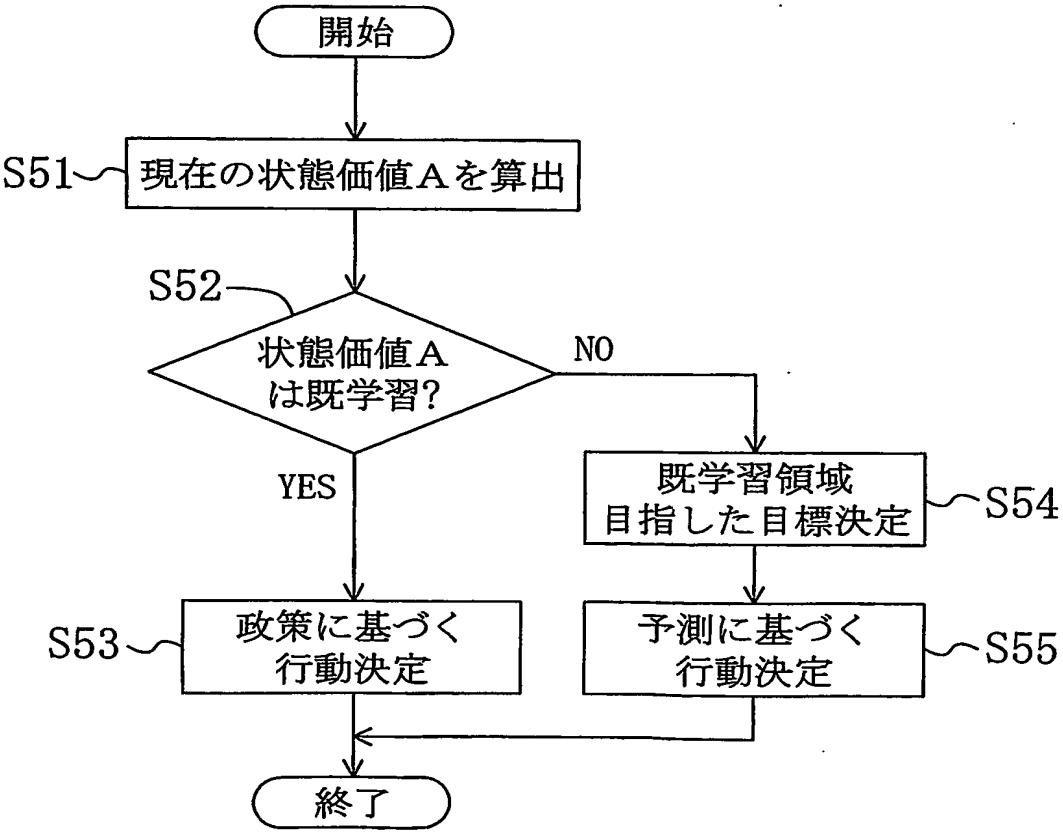
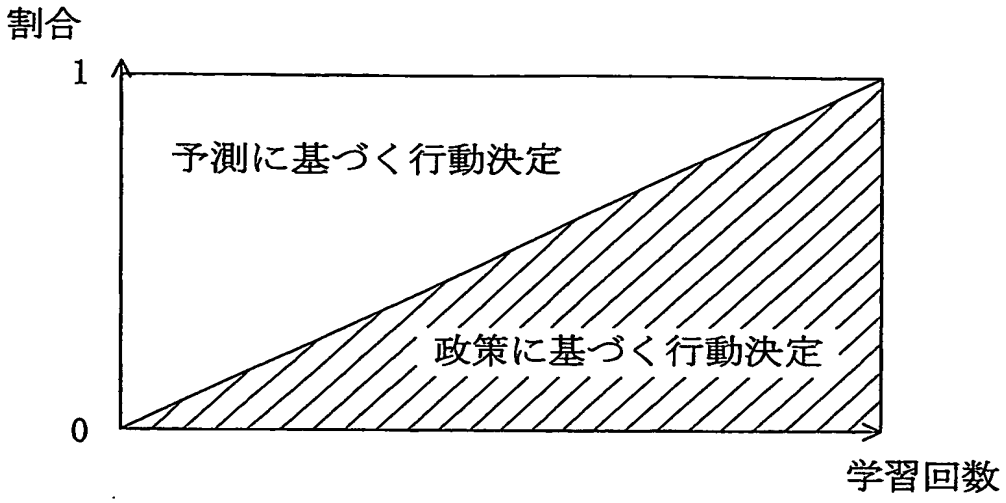

開始
S51
現在の状態価値Aを算出
S52
状態価値A
は既学習?
NO
YES
既学習領域
目指した目標決定
S54
S53
政策に基づく
行動決定
予測に基づく
行動決定
S55
終了

FIG. 19



INTERNATIONAL SEARCH REPORT

International application No.
PCT/JP2004/000971

A. CLASSIFICATION OF SUBJECT MATTER
Int.Cl[7] G06N5/00

According to International Patent Classification (IPC) or to both national classification and IPC

B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)
Int.Cl[7] G06N5/00

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched

Jitsuyo Shinan Koho	1922-1996	Toroku Jitsuyo Shinan Koho	1994-2004
Kokai Jitsuyo Shinan Koho	1971-2004	Jitsuyo Shinan Toroku Koho	1996-2004

Electronic data base consulted during the international search (name of data base and, where practicable, search terms used)
JOIS, DAITO MASARU+OMORI TAKASHI+MORIKAWA KOJI+OKA NATSUKI
(in Japanese)

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
X	JP 9-297690 A (Nippon Telegraph And Telephone Corp.), 18 November, 1997 (18.11.97), Column 2, line 35 to column 3, line 39; Fig. 9 (Family: none)	1,3,12,14
Y	same as the above	4,7-10,15
A	same as the above	2,5,6,11,13
Y	Taku YOSHIOKA et al., "Kyoka Gakushu ni yoru Othello no Senryaku Kakutoku", Dai 32 Kai Jinko Chino Kisoron Kenkyukai Shiryo, The Japanese Society for Artificial Intelligence, Japan, 26 March, 1998 (26.03.98), pages 115 to 120; page 116, left column, lines 22 to 35	4,7,8,15

☒ Further documents are listed in the continuation of Box C. ☐ See patent family annex.

* Special categories of cited documents:
"A" document defining the general state of the art which is not considered to be of particular relevance
"E" earlier application or patent but published on or after the international filing date
"L" document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)
"O" document referring to an oral disclosure, use, exhibition or other means
"P" document published prior to the international filing date but later than the priority date claimed
"T" later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention
"X" document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone
"Y" document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art
"&" document member of the same patent family

Date of the actual completion of the international search 24 March, 2004 (24.03.04)	Date of mailing of the international search report 06 April, 2004 (06.04.04)
Name and mailing address of the ISA/ Japanese Patent Office Facsimile No.	Authorized officer Telephone No.

Form PCT/ISA/210 (second sheet) (January 2004)

INTERNATIONAL SEARCH REPORT

International application No.
PCT/JP2004/000971

C (Continuation). DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
Y	JP 9-245015 A (Toshiba Corp.), 19 September, 1997 (19.09.97), Column 20, lines 19 to 26; Fig. 6 (Family: none)	9,10
Y	Takeshi ITO, "Shogi no Kansosen ni Mirareru Tansaku no Henka Katei", Information Processing Society of Japan Kenkyu Hokoku, Information Processing Society of Japan, Vol.2001, No.28, 16 March, 2001 (16.03.01), pages 47 to 54; page 48, left column, lines 31 to 40	10
A	Yasuharu KOIKE, Kenji Dotani, "Multi Step Jotai Yosoku o Mochiita Kyoka Gakushu-Chumokuten Idoten o Koryo shita Jidosha Unten Model", The Institute of Electronics, Information and Communication Engineers Gijutsu Kenkyu Hokoku, The Institute of Electronics, Information and Communication Engineers, Vol.98, No.673, 18 March, 1999 (18.03.99), pages 241 to 247	1-15
A	Koichiro TAKITA et al., "Pulse Neural Network ni Okeru Network Kakuchogata Kyoka Gakushu Algorithm", The Institute of Electronics, Information and Communication Engineers Gijutsu Kenkyu Hokoku, The Institute of Electronics, Information and Communication Engineers, Vol.99, No.684, 13 March, 2000 (13.03.00), pages 181 to 188, full text; all drawings	1-15
A	Shingo OCHIAI et al., "Kyoka Gakushu o Mochiita Kyugi Dosa no Yosoku to Seigyo", The Institute of Electronics, Information and Communication Engineers Gijutsu Kenkyu Hokoku, The Institute of Electronics, Information and Communication Engineers, Vol.97, No.69, 23 May, 1997 (23.05.97), pages 69 to 75; full text; all dawings	1-15

国際調査報告

国際出願番号　PCT／JP2004／000971

A.　発明の属する分野の分類（国際特許分類（IPC））
Int. Cl⁷ G06N5／00

B.　調査を行った分野
調査を行った最小限資料（国際特許分類（IPC））
Int. Cl⁷ G06N5／00

最小限資料以外の資料で調査を行った分野に含まれるもの
日本国実用新案公報　1922－1996年
日本国公開実用新案公報　1971－2004年
日本国登録実用新案公報　1994－2004年
日本国実用新案登録公報　1996－2004年

国際調査で使用した電子データベース（データベースの名称、調査に使用した用語）
JOIS，大東優+大森隆司+森川幸治+岡夏樹

C.　関連すると認められる文献

引用文献のカテゴリー＊	引用文献名　及び一部の箇所が関連するときは、その関連する箇所の表示	関連する請求の範囲の番号
	JP　9－297690　A　（日本電信電話株式会社） 1997．11．18	
X	第2欄第35行－第3欄第39行，図9 （ファミリーなし）	1，3， 12，14
Y	同上	4，7－10， 15，

☒ C欄の続きにも文献が列挙されている。　☐ パテントファミリーに関する別紙を参照。

＊ 引用文献のカテゴリー
「A」特に関連のある文献ではなく、一般的技術水準を示すもの
「E」国際出願日前の出願または特許であるが、国際出願日以後に公表されたもの
「L」優先権主張に疑義を提起する文献又は他の文献の発行日若しくは他の特別な理由を確立するために引用する文献（理由を付す）
「O」口頭による開示、使用、展示等に言及する文献
「P」国際出願日前で、かつ優先権の主張の基礎となる出願の日の後に公表された文献
「T」国際出願日又は優先日後に公表された文献であって出願と矛盾するものではなく、発明の原理又は理論の理解のために引用するもの
「X」特に関連のある文献であって、当該文献のみで発明の新規性又は進歩性がないと考えられるもの
「Y」特に関連のある文献であって、当該文献と他の1以上の文献との、当業者にとって自明である組合せによって進歩性がないと考えられるもの
「&」同一パテントファミリー文献

国際調査を完了した日 24．03．2004	国際調査報告の発送日 06．4．2004
国際調査機関の名称及びあて先 日本国特許庁（ISA／JP） 郵便番号100－8915 東京都千代田区霞が関三丁目4番3号	特許庁審査官（権限のある職員） 漆原 孝治　5B　3145 電話番号　03－3581－1101　内線　3546

様式PCT／ISA／210（第2ページ）（2004年1月）

国際調査報告

国際出願番号　PCT／JP2004／000971

C（続き）．　関連すると認められる文献

引用文献のカテゴリー＊	引用文献名　及び一部の箇所が関連するときは、その関連する箇所の表示	関連する請求の範囲の番号
A	同上	2，5，6，11，13
	吉岡琢，外２名著，“強化学習によるオセロの戦略獲得”，第３２回人工知能基礎論研究会資料，社団法人人工知能学会，日本，１９９８．０３．２６，第１１５－１２０頁	
Y	第１１６頁左欄２２－３５行	4，7，8，15
	ＪＰ　９－２４５０１５　Ａ　（株式会社東芝）１９９７．０９．１９	
Y	第２０欄第１９－２６行，図６（ファミリーなし）	9，10
	伊藤毅志 著，“将棋の感想戦に見られる探索の変化過程”，情報処理学会研究報告，社団法人情報処理学会，第２００１巻，第２８号，２００１．０３．１６，第４７－５４頁	
Y	第４８頁左欄第３１－４０行	10
	小池康晴・銅谷賢治共著，“マルチステップ状態予測を用いた強化学習－注目点移動点を考慮した自動車運転モデル”，電子情報通信学会技術研究報告，社団法人電子情報通信学会，第９８巻，第６７３号，１９９９．０３．１８，第２４１－２４７頁	
A	全文，全図	1－15
	瀧田航一朗，外２名著，“パルスニューラルネットワークにおけるネットワーク拡張型強化学習アルゴリズム”，電子情報通信学会技術研究報告，社団法人電子情報通信学会，第９９巻，第６８４号，２０００．０３．１３，第１８１－１８８頁	
A	全文，全図	1－15
	落合真吾，外２名著，“強化学習を用いた球技動作の予測と制御”，電子情報通信学会技術研究報告，社団法人電子情報通信学会，第９７巻，第６９号，１９９７．０５．２３，第６９－７５頁	
A	全文，全図	1－15